The genomic and transcriptomic landscape of clinical *Escherichia coli* and *Pseudomonas aeruginosa* isolates

Von der Fakultät für Lebenswissenschaften

der Technischen Universität Carolo-Wilhelmina zu Braunschweig

zur Erlangung des Grades eines

Doktors der Naturwissenschaften

(Dr. rer. nat.)

genehmigte

D i s s e r t a t i o n

von Uthayakumar Muthukumarasamy
aus Manalmedu, Tamilnadu, India

1. Referent: Professor Dr. Karsten Hiller
2. Referentin: Professorin Dr. Susanne Haeussler
eingereicht am: 05.12.2018
mündliche Prüfung (Disputation) am: 14.03.2019

Druckjahr 2019

Vorveröffentlichungen der Dissertation

Teilergebnisse aus dieser Arbeit wurden mit Genehmigung der Fakultät für Lebenswissenschaften, vertreten durch den Mentor der Arbeit, in folgenden Beiträgen vorab veröffentlicht:

Publikationen

Bielecki, P., Muthukumarasamy, U., Eckweiler, D., Bielecka, A., Pohl, S., Schanz, A., Niemeyer, U., Oumeraci, T., von Neuhoff, N., Ghigo, J.M. and Häussler, S., 2014. In vivo mRNA profiling of uropathogenic Escherichia coli from diverse phylogroups reveals common and group-specific gene expression profiles. *MBio*, *5*(4), pp.e01075-14.

Bruchmann, S., Muthukumarasamy, U., Pohl, S., Preusse, M., Bielecka, A., Nicolai, T., Hamann, I., Hillert, R., Kola, A., Gastmeier, P. and Eckweiler, D., and Häussler, S., 2015. Deep transcriptome profiling of clinical Klebsiella pneumoniae isolates reveals strain and sequence type-specific adaptation. *Environmental microbiology*, 17(11), pp.4690-4710.

Hornischer, K., Khaledi, A., Pohl, S., Schniederjans, M., Pezoldt, L., Casilag, F., Muthukumarasamy, U., Bruchmann, S., Thöming, J., Kordes, A. and Häussler, S., 2018. BACTOME - a reference database to explore the sequence-and gene expression-variation landscape of Pseudomonas aeruginosa clinical isolates. *Nucleic acids research*, gky895, 2018 Oct 1.

To,

My mom,

who, for the love of me,

never questioned my quest for science,

and so taught me how to love,

and

My dad,

who, for the love of me,

always questioned my quest for science,

and so taught me how to attain knowledge.

காக்கைக்கா காகூகை கூகைக்கா காகாக்கை

கோக்குக்கூ காக்கைக்குக் கொக்கொக்க – கைக்கைக்குக்

காக்கைக்குக் கைக்கைக்கா கா.

-காளமேகப் புலவர்

Crow cannot defeat an owl at night; and an owl cannot defeat a crow during the day. To properly rule a kingdom, the king should wait patiently for the right opportunities and use them, like a crane waiting patiently in the water for fish. Otherwise protection from enemies will be beyond the reach of even powerful kingdoms.

-Kalamegam, ca. 15th century.

On the virtue of patience and waiting for the right opportunity.

(Written with only one consonant, k.)

"Life would not long remain possible in the absence of microbes."

-Louis Pasteur, ca. 1880

Zusammenfassung

Schnell voranschreitende Entwicklungen auf dem Gebiet der DNA-Sequenziertechnologien ermöglichen die Erzeugung immer größerer Datenmengen. Mit der Entwicklung effizienter bioinformatischer Plattformen eröffnen diese großen Mengen genomischer Daten eine Vielzahl von Möglichkeiten, unser Wissen über die genetische Zusammensetzung und Evolution einzelner Spezies zu erweitern. Der Schwerpunkt dieser Arbeit besteht in der Entwicklung geeigneter computergestützter Prozesse, um relevante Informationen aus Transkriptom- und Genomdaten zu filtern und wichtige biologische Fragestellungen zu beantworten.

Die Ziele der Arbeit waren dabei in zwei Bereiche gegliedert: i) die Entwicklung einer pangenom-basierten Analysemethode von RNA-Sequenzierungsdaten, um *ex vivo* Genexpressionsprofile von uropathogenen *Escherichia coli* Isolaten zu analysieren und ii) die Erstellung der Konsensus-Sequenz des *Pseudomonas aeruginosa* Kern-Genoms, um in klinischen Isolaten von *P. aeruginosa* Einzelnukleotidmutationen (single nucleotide polymorphisms, SNPs) mit hoher Genauigkeit zu erkennen und patho-adaptive Mutationen zu identifizieren.

Hierfür wurde das *E. coli* Pangenom erstellt und verwendet, um RNA-Sequenzierungsdaten von diversen klinischen *E. coli* Isolaten, die von einer akuten Harnwegsinfektionen stammen, zu analysieren. Während das *in vivo* Genexpressionsprofil innerhalb der untersuchten 21 *E. coli* Isolate zu einem großen Teil konserviert war, konnten dennoch erhebliche Unterschiede in der Expression des akzessorischen Genoms beobachtet werden, die sich auch in der phylogenetischen Verwandtschaft der Stämme widerspiegelte. Des Weiteren wurden in den *in vivo* Expressionsprofilen im Unterschied zu *in vitro* Transkriptomen unter anderem Virulenzfaktoren und kleine regulatorische RNAs (sRNAs) gefunden, sowie weitere Gene, die bisher nicht mit bakterieller Virulenz assoziiert wurden.

Neben der pangenom-basierten RNA-Sequenzierungsanalyse von klinischen *E. coli* Isolaten wurde die genetische Variabilität innerhalb der Genome von 99 klinischen *P. aeruginosa* Isolaten näher untersucht. Für jedes Kern-Gen wurde die Konsensus-Sequenz basierend auf der berechneten Anordnung des am häufigsten vorkommenden Nukleotids erstellt, um es als Referenz für die Identifikation der SNPs in allen klinischen Isolaten zu verwenden. Dabei

wurden die gefundenen SNPs unterschiedlichen Gruppen zugeordnet, je nachdem, ob sie Ursprung des klonalen Hintergrundes, nur in einem Isolat vorkommend oder für verschiedene phylogenetische Untergruppen charakteristisch waren (phylogenetisch unabhängig). Während die Mehrzahl der SNPs in eine der ersten beiden Gruppen fielen, konnten 2.252 Gene mit einer oder mehr nicht-synonymen Mutationen identifiziert werden, die in die Gruppe der phylogenetisch unabhängigen SNPs fallen. Des Weiteren ergab die Analyse des Verhältnisses von nicht-synonymen (dN) zu synonymen Substitutionen (dS) dN/dS bei 3.814 Genen, dass das Kern-Genom nicht unter Selektionsdruck steht. Basierend auf dieser Analyse können in Zukunft patho-adaptive Mutationen des akzessorischen Genoms identifiziert werden, sowie die vollständige genetische Variation und adaptive Evolution der klinischen *P. aeruginosa* Isolate beschrieben werden.

Zusammengefasst untersucht diese Arbeit die Möglichkeiten pangenom-basierter Ansätze bei der Analyse von Transkriptom- und Genomdaten von klinischen *E. coli* und *P. aeruginosa* Isolaten. Die daraus gewonnen Erkenntnisse konnten bereits entscheidend zum Verständnis der Sequenz-Variationen, die in der Umwelt und im menschlichen Körper selektiert werden und zu Adaptation und Pathogenität führen, beitragen. Zukünftig können die zu Grunde liegenden Methoden genutzt werden, um die Evolutionsprozesse in pathogenen Populationen zu verstehen. Dies ist vor allem wichtig, um den Zusammenhang zwischen Diversität, der Struktur bakterieller Gemeinschaften und ihrer Funktion zu verstehen.

Summary

Large amounts of genomic data have been obtained due to the rapid advances in DNA sequencing technology. With efficient computational platforms, these genomic data can provide many possibilities to improve our knowledge on species evolution and their genetic makeup. The general interest of this thesis is to facilitate studies on important biological questions by attaining the relevant information from transcriptomic and genomic data.

The aims of my thesis were i) to develop the pan-genome based RNA sequencing data analysis pipeline in order to analyze *ex vivo* gene expression profiles of uro-pathogenic *Escherichia coli* isolates and ii) to create the consensus sequence of the *Pseudomonas aeruginosa* core genome in order to identify single nucleotide polymorphisms (SNPs) at high accuracy and to find the patho-adaptive mutations in *P. aeruginosa* clinical isolates.

To address these aims I developed and used the pan-genome of *E. coli* in order to map and analyze the RNA sequencing reads obtained from a diverse array of clinical *E. coli* isolates that were associated with a clinical course of an acute urinary tract infection. Whereas the *in vivo* gene expression profiles of the majority of genes were conserved among the 21 *E. coli* strains, the specific gene expression profiles of the accessory genome were diverse and reflected phylogenetic relationships. Furthermore, genes transcribed *in vivo* relative to laboratory media included well-described virulence factors, small regulatory RNAs, as well as genes not previously linked to bacterial virulence.

In addition to the pan-genome based RNA sequencing data analysis of clinical *E. coli* isolates, whole genome sequencing data was used to gain insights into the genetic variations of 99 clinical *P. aeruginosa* isolates. I created the consensus sequence for every core gene based on the calculated order of the most frequent nucleotide. I used it as reference for the identification of SNPs in all clinical isolates. The identified SNPs were classified into clonal-specific SNPs, single SNPs (occurring only in one isolate) and phylogenetically independent SNPs (inter clonal SNPs). The majority of the SNPs were clonal-dependent and single SNPs. However, I identified a large set of 2,252 genes which had one or more phylogenetically independent non-synonymous mutation. Moreover, the analysis of the ratio of nonsynonymous substitutions (dN) to synonymous substitutions (dS), dN/dS on 3,814 genes revealed that the core genome is not under

selection pressure. I provide a framework so that in the future the pipeline can be used to also find the patho-adaptive mutations from the accessory genome as well as to describe the complete genetic variations and adaptive evolution of the clinical *P. aeruginosa* isolates.

In summary, this thesis explores pan-genome-based as well as consensus sequence-based approaches on transcriptomic and genomic sequencing data of clinical isolates of *E. coli* and *P. aeruginosa* respectively. The results of the thesis contributed to understanding of sequence variations that are selected in the environment of the human host and lead to bacterial adaptation and pathogenicity. In the future the developed analytical tools and the concepts will be exploited for studies that aim for understanding the evolutionary processes in pathogenic populations. This is not only important for the basic scientific research, but also to understand the link between diversity and community structure and function.

Contents

List of Figures

List of Tables

List of Abbreviations

BLAST	basic local alignment search tool
BLASTN	basic local alignment search tool nucleotide
BLASTP	basic local alignment search tool protein
CDS	coding sequence
cgMLST	core genome multi locus sequence typing
DDBJ	DNA data bank of japan
dN/dS	the ratio of nonsynonymous to synonymous SNP rates
E. coil	*Escherichia coli*
EMBL	european molecular biology laboratory
GO	gene ontology
HMP	human microbiome project
indels	insertions / deletions
INSDC	international nucleotide sequence database collaboration
K. pneumoniae	*Klebsiella pneumoniae*
KEGG	kyoto encyclopedia of genes and genomes
MLST	multi locus sequence typing
NCBI	national center for biotechnology information
ncRNAs	non-coding RNAs
NGS	Next-generation sequencing
nRPK	normalized reads per kilobase
OGs	orthologous group
P. aeruginosa	*Pseudomonas aeruginosa*
PseudoCAP	Pseudomonas community annotation project
rRNAs	ribosomal RNAs
RNA-Seq	RNA sequencing
RPG	read count per gene
STs	sequence types
SNPs	single nucleotide polymorphisms

sRNAs	small RNAs
TTSS/T3SS	type three secretion system
UTI	urinary tract infection
UPEC	uropathogenic *Escherichia coli*

Glossary of terms

Big data	A vast, complex and rapidly growing data (in volume, variety and velocity) that require a high-end computing and analytical frameworks for its ample value
Pan-genome	A collection of non-redundant gene repertoire that is present in a given dataset of 'n' genomes. It can be categorized into core genome, accessory or flexible genome, and singletons (strain/species specific genes)
Core genome	Genes that are present in all the 'n' genomes of given dataset
Accessory or flexible genome	Genes that are present in more than one genome but not in all genomes of given dataset
Singletons	Genes that are specific to a particular genome (unique)
"Soft" core genes	Genes that are seldom absent in few, yet present in most of the given genomes
Non-redundant gene repertoire	A collection of representative genes from every distinct gene group of 'n' genomes
Reciprocal hits	A relationship between the query sequence and the best-hit sequence is always similar and interchangeable (a bidirectional orthologous gene)
Reciprocal clade	The relationship among a group of genes is always similar and interchangeable within that particular group
cgMLST	Core genome (based) Multi Locus Sequence Typing
Consensus sequence	A reference sequence derived from the group of orthologous gene sequences based on the most frequently occurring nucleotide at each position. Of note, the orthologous gene sequences are identical in length.
Patho-adaptive mutation (adaptive pathogenicity)	Mutations indicate a genetic mechanism for enhancing bacterial virulence without horizontal transfer of specific virulence factors. Those mutations confer a strong selective advantage over the bacterial clone in the virulence niche. Adapted from [22]

Introduction

Introduction

Nowadays, big data have a pronounced impact in most of the scientific disciplines. After the completion of human reference genome in 2003, the sequencing technologies have revolutionized the modern biological sciences, particularly in microbial genomics, which has led to enhance our understanding on how genome variation impact health and disease. Consequently, the fields of omics especially genomics, transcriptomics and proteomics have produced a massive amount of data which requires computationally sophisticated platforms for their analyses and interpretations [1]. Simultaneously, the volume of sequencing data in the public databases is doubling approximately in every 18 months and therefore uncovering the accumulated information on these databases needs to be processed rapidly to gain improved knowledge. For instance, although hundreds of genomes and/or transcriptomes have been obtained from many individual species, the data analysis pipelines still utilize one of the genomes as a reference to study their genotypes and/or phenotypes. This clearly implicates the loss of information on several of the acquired genes and may therefore have substantial consequences on subsequent interpretations. Taken together, it is essential to have an analytical framework based on multiple genomes of a species than a single reference genome so that the maximum number of sequenced reads can be overlaid during the downstream analysis. This approach will be useful for global transcriptome and genome landscape studies to delineate the genotypes and increasing number of the complex phenotypes.

1.1 Bioinformatics, DNA research and Microbial Genomics: A historical perspective

During the last half century, there were many breakthroughs noticed in the field of DNA research and bioinformatics [2]. The phenomenal breakthrough in molecular biology research was the discovery of the double helical structure of the DNA molecule by Francis Crick and James Watson in 1953 [3]. The realm of sequencing had started to grow when Frederick Sanger developed a technique to sequence the DNA [4]. Subsequently, the idea of sequencing and analyzing the whole human genome was proposed with the automation of DNA sequencing.

Later, subordinated by two major objectives; first, sequencing the model organisms such as yeast, fly, nematode, bacteria, mouse, and a flowering plant in order to understand the biological process in particular genetics; second, to develop novel tools and technologies to accelerate the genomic research. After fifteen years of devoted research, the community obtained the first finished human reference genome officially in 2003 while obtaining the genomes of other model organisms, that have immensely transformed our understanding of the genetic makeup of an individual organism [5].

In the interim, computational biology was gently evolved when Margaret Oakley Dayhoff developed the atlas of protein in 1965 that is considered as the founding text for bioinformatics [6]. Afterwards, a method based on mutation distances was developed to construct the phylogenetic tree in 1967 [7]. Further, one of the first applications of dynamic programming to align and find the global similarities in the biological sequences was developed [8], which is popularly known as a Needleman-Wunch algorithm or global alignment. It remains as a basis for developing another major algorithm, known as Smith-Waterman (or local alignment) algorithm, to find the similar regions between the sequences, despite entirety [9]. This is followed by the most ever successor algorithm, called BLAST, Basic Local Alignment Search Tool [10] which is still considered as the people's choice to find the similarities of the given sequence. A family of BLAST algorithms was developed sequentially with a focus of different sequence entities to study about orthologous genes, their functions, phylogeny and evolution [11].

The major paradigm shift in the field of microbiology had begun when the first complete bacterial genome *Haemophilus influenza* was sequenced in 1995 [12]. Later, many bacterial genomes, including *Escherichia coli* completely sequenced in a short span of time [13]. However, the revolutionary next generation sequencing technologies has made it possible to sequence hundreds of thousands of genomes, more quickly and cost-effectively, which led not only to the large scale projects such as 1000 genome project and Human Microbiome Project (HMP) to understand human health and disease, but also to an individual research group that study more on hospital settings and ecological perspectives. As a consequence, the scientific community has generated vast, diverse and complex genomes to transform its value for future application. Several methods, tools and databases have been developed, ranging from the de

Bruijn Graph to hidden markov model, assembly to annotation and read alignment to variant calling, to make use of such overflowing data to understand how genomic variations control health and disease. Also, gene ontology (GO) consortium and the derived public databases like KEGG play a major role in functional and signaling pathway related studies [14, 15]. Recently, a single study identified more than 15% of the bacterial domain, which includes greater than 35 phyla, and is referred as candidate phyla radiation (CPR) [16]. This is the only study indicating the fact that the community will witness many more novel genomes in the near future.

The robust coalescence of bacterial genome sequencing and bioinformatics has driven a revolution in data analytics and augmented our understanding in bacterial function, evolution, interactions within themselves, with their niches, and with their hosts and at the same time expanding several horizons for translational medicine. As an example, genomic variations and their complex phenotypes of three different strains of *Pseudomonas aeruginosa* were revealed by sequencing and analysis. The first largest complete bacterial genome sequence of *P. aeruginosa* PAO1 strain was sequenced in 2000 that reflected their evolutionary adaptations in different niches and the effects of antimicrobial resistance of the species [17]. After few years, another major strain of *P. aeruginosa* PA14 was sequenced. This study has revealed that the virulence is due to both multifactorial and combinatorial in *P. aeruginosa* in different genetic backgrounds [18]. The third strain PA7 of *P. aeruginosa* was sequenced in 2010, which was identified as a taxonomic outlier of the species and for their multi resistant mechanisms [19]. Thus, exploring the complex and diverse genomes of individual species aids to describe its genomic landscape that plays a crucial role in adaptation strategies and pathogenicity. Further, *Pseudomonas aeruginosa* genome database was developed with a major focus to facilitate community based (updated) genome annotation (referred as PseudoCAP) [20]. With efficient bioinformatics platforms, these genomic data can provide many possibilities to improve our knowledge of species evolution and their genetic makeup. Some of the major breakthroughs in both bioinformatics and sequencing domain (in general) are represented as a timeline in Figure 1, with additional information on *P. aeruginosa* and a few other species as interesting cases [2].

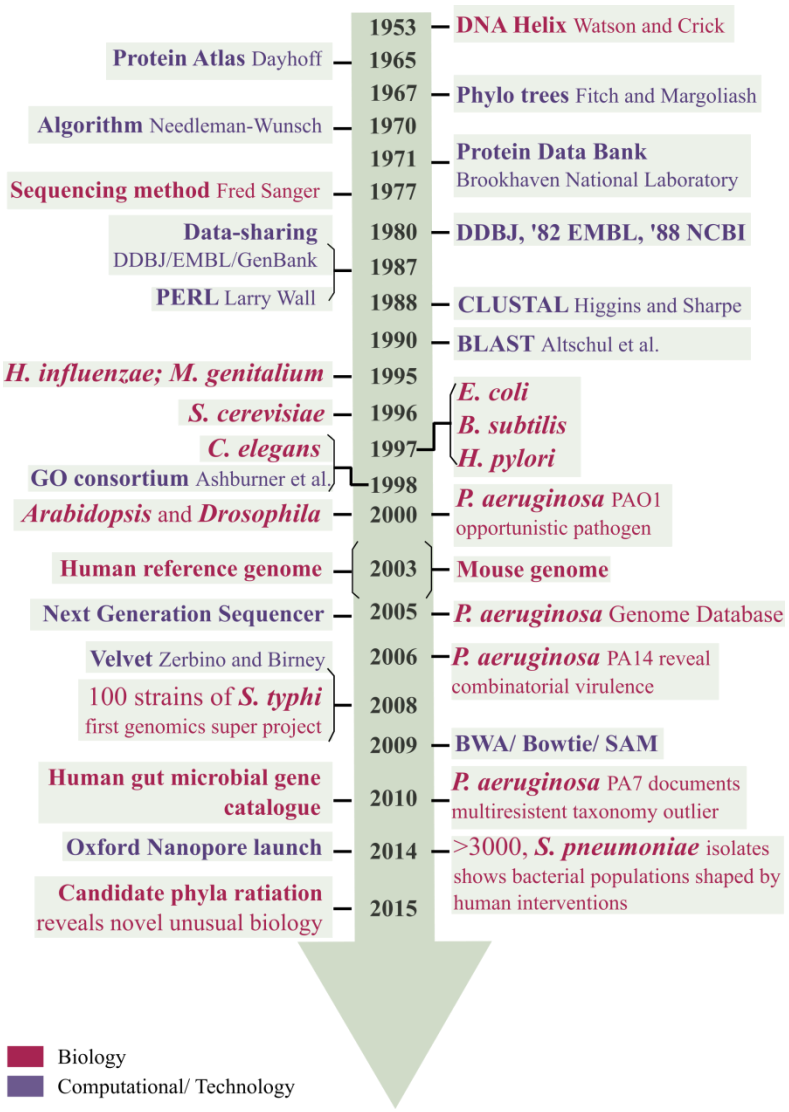


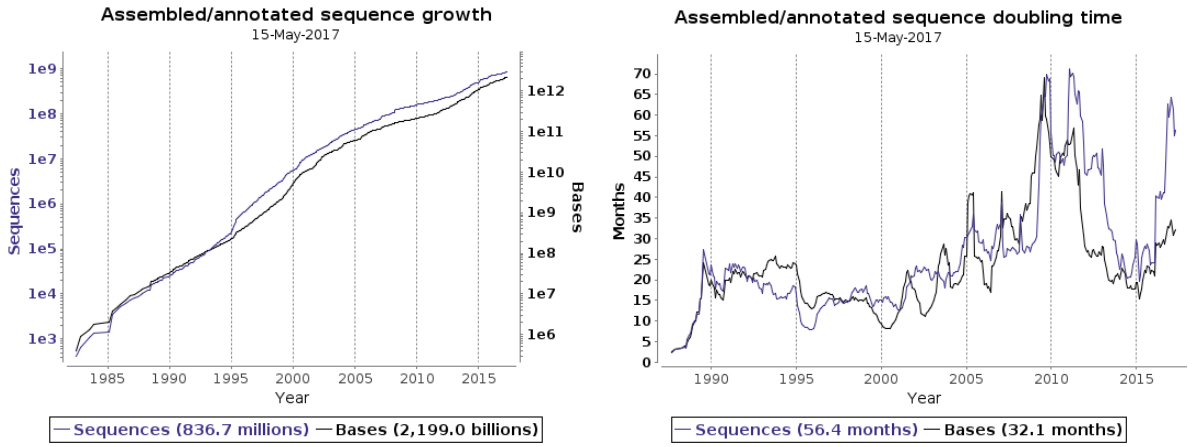
Figure 1: Some of the major milestones in bioinformatics and genomic sequencing.

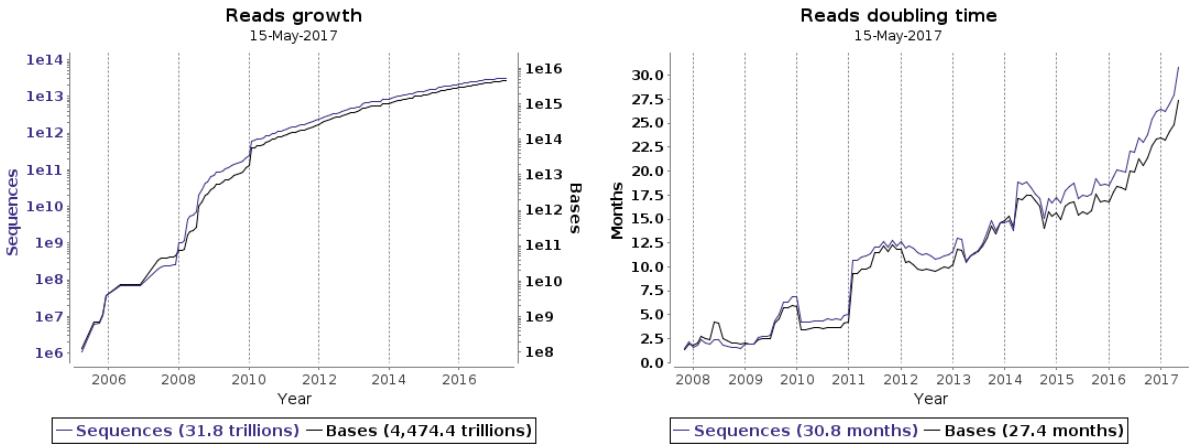
H. influenza, Haemophilus influenza; M. genitalium, Mycoplasma genitalium; S. cerevisiae, Saccharomyces cerevisiae; E. coli, Escherichia coli; B. subtilis, Bacillus subtilis; H. pylori, Helicobacter pylori; C. elegans, Caenorhabditis elegans; P. aeruginosa; Pseudomonas aeruginosa (strains - PAO1, PA14, PA7)*; S. typhi, Salmonella enterica subsp. enterica serovar typhi; S. pneumoniae, Streptococcus pneumoniae.*

1.2 Growth of data in the public genome databases

The recent advancements in the next generation sequencing technologies have provided most comprehensive genomic and transcriptomic information. The sequencing data has been deposited in one of the public genome databases, namely, Genbank of NCBI (National Center for Biological Information) or ENA of EMBL (European Nucleotide Archive at European Molecular Biology Laboratory) or DDBJ of NIG (DNA Data Bank of Japan at National Institute of Genetics) grew at unprecedented speed. They are the members of INSDC (International Nucleotide Sequence Database Collection) which exchange their data on a daily basis. So these databases reflect the same information at any given time [21-24]. Over the last decade, the data acquisition has steadily increased and currently storing approximately 830 million processed data and about 30 trillion read data. Hence, these databases represent largest in the world, primarily, not only for the storage and retrieval but also for the processing of any derived data. The approximate doubling time of data growth in these databases is predicted to be 18 months, according to the Moore's law, although Illumina and other study estimated the doubling time to be 12 months and 7 months, respectively [25]. Growth of the processed data and the read data (with their doubling time) is presented in Figure 2.

(a) (b)





(c) (d)

Figure 2: Data growth in the genome databases Genbank/EMBL/DDBJ.
The graphs illustrate the growth of genomic data in the public genome databases between 1982 to present. (a) Assembled and annotated data in terms of number of bases and number of sequences, that have steadily risen from 680000 to 2 trillion and 606 to 837 million respectively (b) with an average doubling time of 34 months. (c) Next generation sequencing read data between 2005 to present, shows the rapid growth of about 32 trillion reads and 4,474 trillion bases (d) with an average doubling time of 27 months. [Adapted from ENA statistics as of 15th May 2017]

Until now, the public genome databases, such as Genbank, possess twenty divisions which include 12 taxonomical, 5 high-throughputs, 1 patent, 1 transcriptomic and 1 whole genome shotgun division that makes the data retrieval easier [24]. The sequence records are allocated in one of twenty divisions based on either taxonomy or the sequencing strategy used to obtain the data. Currently, the amount of genomic data is massively expanding followed by the transcriptomic data and this type of division-wide growth is presented in Figure 3.

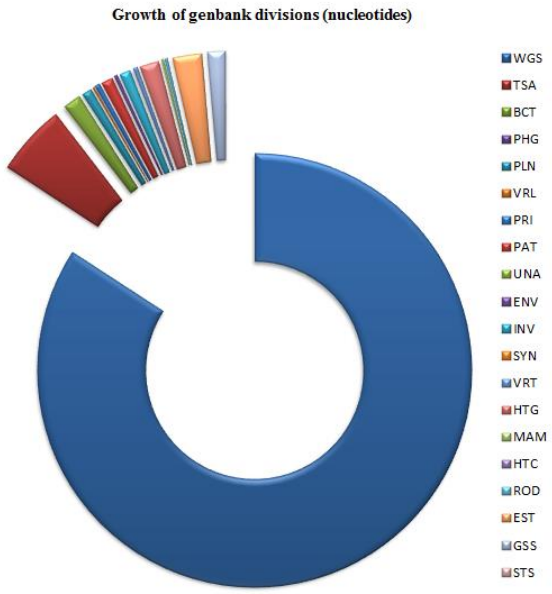


Figure 3: Growth of Genbank divisions.

Whole genome sequencing (WGS) data tops the database with its size about 1.6 trillion, followed by transciptome shotgun data (TSA) data which is greater than 100 billion (as of the Genbank 215 release, august 2016) [24].

(WGS, Whole genome shotgun data; TSA, Transcriptome shotgun data; BCT, Bacteria; PHG, Phages; PLN, Plants; VRL, Viruses; PRI, Primates; PAT, Patent sequences; UNA, Unannotated; ENV, Environmental samples; INV, Invertebrates; SYN, Synthetic; VRT, Other vertebrates; HTG, High-throughput genomic; MAM, Other mammals; HTC, High-throughput cDNA; ROD, Rodents; EST, Expressed sequence tags; GSS, Genome survey sequences; STS, Sequence tagged sites)

The genome databases provide an efficient access over the internet at free of cost through FTP (file transfer protocol), web browsers and web services such as SOAP (Simple Object Access Protocol) or REST (Representational state transfer) protocols. Further, Genbank and ENA have many inbuilt tools and platforms to deliver insightful information to all the users [26]. On the other hand, new biological databases are being constantly developed either by using the existing data or coupled with new data or both. The continuous growth of the biological databases, in the last 12 years, is presented in Figure 4. At present, the number of biological databases has reached a peak about 1,685 and plus [27]. They are split into 15 major biological categories and 41 sub-categories. They are also available through web browsers and their descriptions and web-links

are provided in the annual database issue of the journal nucleic acid research (NAR), usually in the month of January, every year.

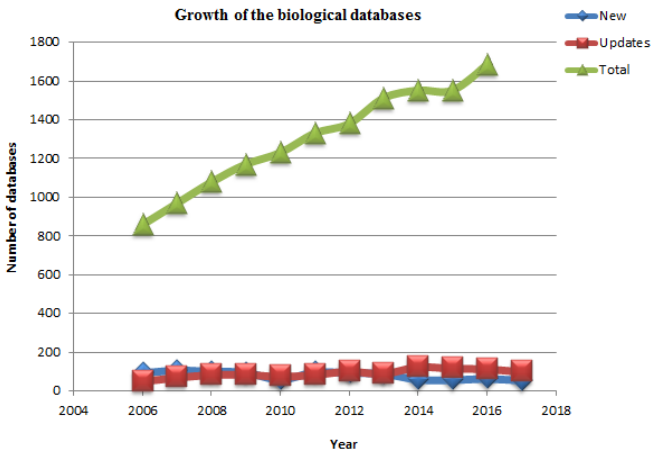


Figure 4: Growth of the biological databases in the last ten years.

The last 12 annual database issues from Nucleic Acid Research journal show that new databases as well as the update of existing databases are frequently developed and published. Presently, the number of biological databases has increased about 1,685 and plus.

Taken together, with the current tools and resourceful platforms, both sequencing as well as obtaining their useful derived data is no longer a hindrance, however, it can take much longer time to devise its biologically relevant information. Hence it compels for new/novel analytical approach with biological criteria, for example to fully understand the genetic make-up of a species.

1.3 Era of big data and omics

Big data is a data resource which is massive in volume, velocity and variety (commonly known as three "V"s) and it requires high performance computing environment and/or high-end analytical frameworks to gain its potential value. Nowadays, Genomics, Astronomy, YouTube and Twitter are the four big data domains in terms of data acquisition, storage, distribution and analysis [25]. The size of the digital universe far exceeds one Zettabyte of data (that is, one billion terabytes or one trillion gigabytes, as of 2012) across all digitized domains and these domains are presented in Figure 5 [adapted from 28]. It implies that both new technologies and

the techniques are required to develop in order to integrate the data that are achieved by different streams to create a descriptive and accurate model, which would eventually aid to identify a global threat at any domain and subsequently to improve our health and life.

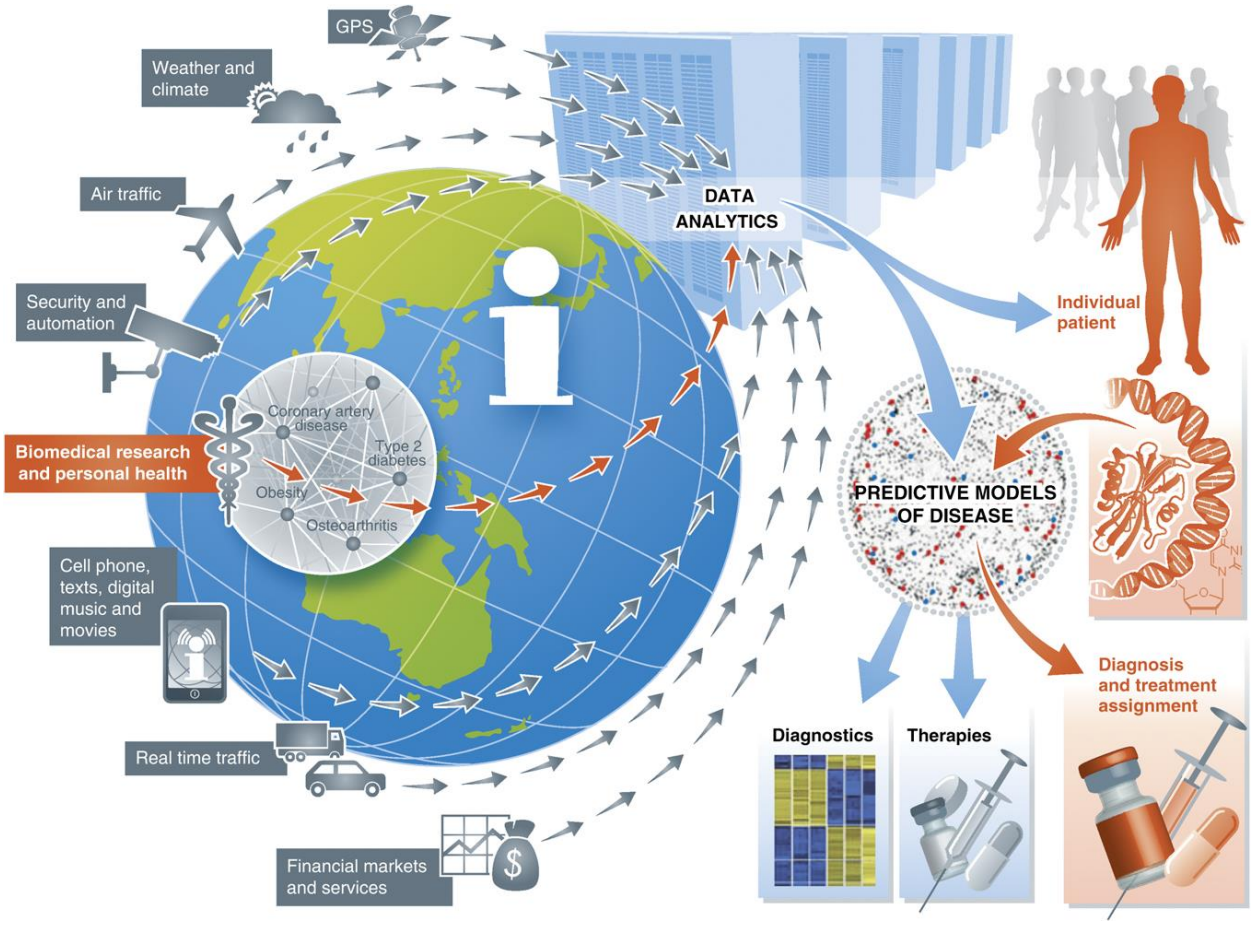


Figure 5: Big data across domains.
Big data is growing at remarkable speed, not only in genomic domain, but across different sectors such as meteorology, telecommunication, security and automation, finance, air traffic, real time traffic. Such overflowing data is strengthening the top-notch data driven approaches to find the next considerable and suspicious threat or improving the diagnosis and treatment of an individual.

With vastly increasing access to bacterial genomes and transcriptomes data, every direction in microbiology has expanded for new insights at unprecedented depth that majorly cover the concept of pan-genome (also known as supra-genome or species-genome). The pan-genome is a collection of non-redundant gene repertoire of the given strains that are usually the representatives of every orthologous gene groups and the unique genes. Further, it is categorized as core genome (equally shared proportion in all the given strains; in other words, genes that are present in all the strains), flexible or accessory genome (shared proportion in two or many of the given strains; in other words, genes that are present in-between 1 to n-1 strains) and unique or singletons that are present only in a specific strain. The first pan-genome based comparative genome analysis was performed in *Streptococcus agalactiae*, identified a significant number of new genes from every individual strain of this species [29]. In the recent years, the concept of pan-genome was extensively applied in more than 50 bacterial species to majorly study its population diversity [30, 31]. However, essential applications of pan-genome are yet to be explored in particular the SNP-identification.

1.4 Need for a pan-genome based analytical framework

Although several methods are developed to find sequence similarities, orthologous identification is still non trivial, especially, either as best reciprocal hits or as best reciprocal clades [32]. The most straightforward approach for identifying orthologous genes is to compare all the genes against all followed by the selection of best hits based on significant pairwise similarities. The cutoff value and the parameters (often an e-value – for sequence identity and sequence length) play a crucial role to identify the orthologous groups (OGs), in both closely related and highly variable strains [33]. Likewise, computational issues (also software choices) in *de novo* genome assembly; gene annotation; read alignment and sequencing error might also influence the total number of OGs. Eventually, the identified orthologous groups relate to the definition of the core genome, for instance, the core genes are generally present in all the given genomes (i.e. 100%) but some might define the core genes even if they are present in greater than 95% of the genomes or more. This clearly indicates that up to 5% of the genomes are being missed. It is even more challenging to develop a reciprocal pan-genome where most of the orthologous groups (OGs) are

the best reciprocal clades (including both core and accessory genes). The reciprocal pan-genome is usually lucid and accurate in read alignment and subset analysis. It is generally able to cover the maximum number of reads than a single reference genome [34]. Further, the unambiguous features (or the subclasses such as core, accessory and singletons) make the downstream analysis easier. The concept of developing a pan-genome is presented in Figure 6.

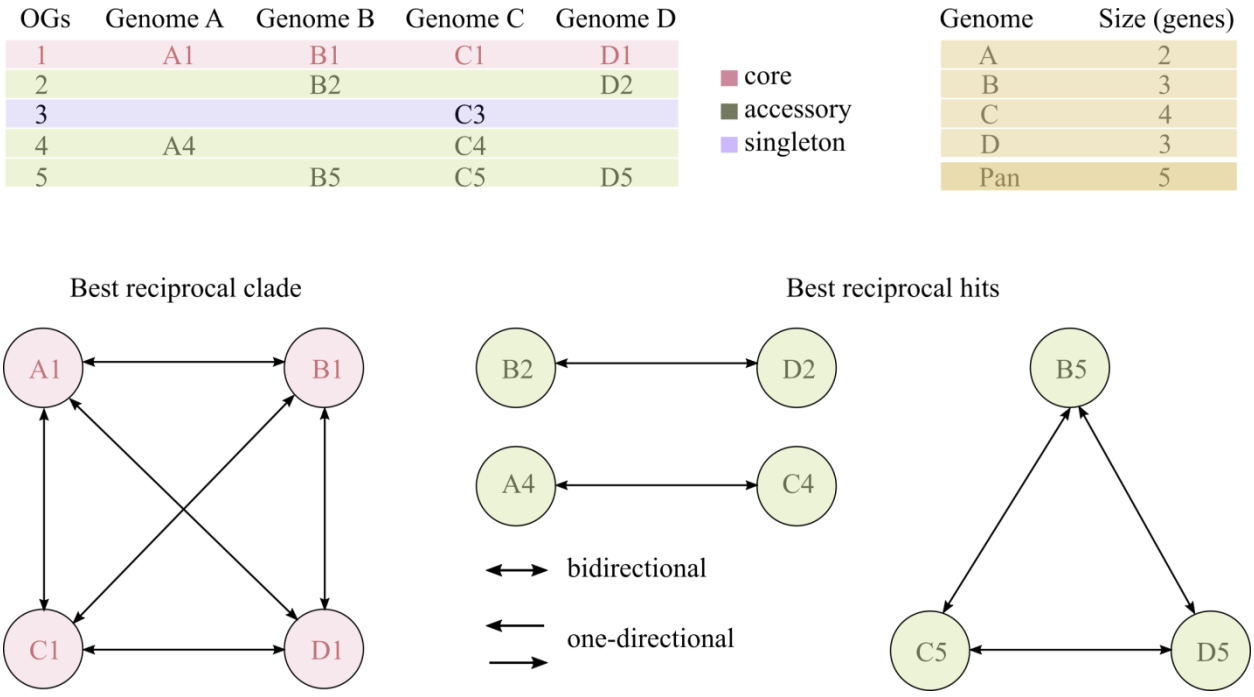


Figure 6: The concept of pan-genome. OGs; Orthologous Groups
Two directional best hits (best reciprocal hits and clades) are mostly detected and if not available, one directional best hits are then included to the gene pool. A total number of representative OGs, which are non-redundant collection of gene repertoire that makes the pan-genome.

At present, core genome can also be used to construct the phylogenetic trees at an unprecedented resolution (cgMLST) which helps to describe the population diversity. Back then, DNA-DNA hybridization method was used to distinguish the species. Later, a multi locus sequence typing (MLST) was most commonly used for the same, though advents in sequencing technology have acknowledged the other markers such as 16S ribosomal RNA that are abundantly present in

microbial and archaeal domains. The common limitation in both MLST and 16s rRNA typing links to the limited genome coverage [35]. A recent genotyping study based on gene-by-gene approach reflects the functional and evolutionary relationships and catalogues bacteria from domain to strain [36]. Thus, core genome based phylogenetic tree shows the high resolution relationship within the group. Genomic coverage of genetic typing methods is shown in Figure 7 [adapted from 36].

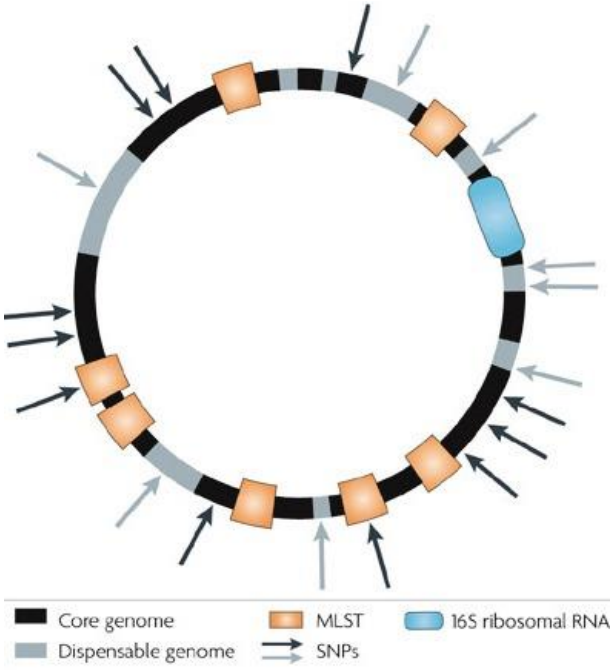


Figure 7: Genomic coverage of genetic typing methods.
Shows core genome in black, accessory genome (dispensable) in grey, MLST regions in orange, 16S ribosomal RNA in blue. SNPs in the core genome are indicated as black arrows whereas SNPs in the accessory genome are indicated as grey arrows.

The pan-genome has become an appropriate reference for sequencing data analysis. It can be used as a reference for read mapping in order to obtain the maximum number of overlaid sequenced reads, both qualitatively and quantitatively. Due to the subclasses (as core, accessory and singletons), pan-genome is abundantly useful in many applications such as typing (cgMLST); phylogenetic relationship within the group/species, i.e. intraclonal or sequence-type

specific relationship; also with others and across the microbial domain, i.e. species and inter-species relationship; describing species diversity by detecting the presence and the absence of genes among the members; functional studies, i.e. exclusive group-specific analysis such as clonal-related or pathogenic members of the group, and most importantly in identifying the true SNPs. It is important to aggregate the strains diversity particularly when analyzing at nucleotide level for identifying the accurate mutations. Many variant callers do not consider this and thereby yielded several strain-dependent SNPs. Because they have been optimized for the diploid genomes but poses difficulties when working with bacterial genomes for variant calling. The most commonly used SAMTools also often yields a high amount of incorrect SNPs when working with the bacterial genomes. In addition to that, a single SNP cut-off value eventually brings the problems of sensitivity (false positives) and specificity (false negatives) to true SNPs [37, 84]. Therefore, pan-genome (at gene level as well as at nucleotide level) based analysis is essential for both genomic and transcriptomic data to uncover the accurate outcome.

1.5 Aims of thesis

To explore a genomic and transcriptomic landscape of a pathogen, it is crucial to study the sequence variations within their strains at both gene and nucleotide level as to fully understand their pathogenicity and the adaptation strategies. The general interest of my thesis is to facilitate studies on important biological questions by attaining the relevant information from transcriptomic and genomic data. The initial motivation for my thesis was to develop improved methods for comparative transcriptomics but later also for comparative genomics, while including multiple references of the same species. I got an opportunity to work on clinical *Escherichia coli, Klebsiella pneumoniae* and *Pseudomonas aeruginosa* isolates.

One of the first aims of this thesis is to develop pan-genome based RNA sequencing data analysis pipeline in order to analyses *ex vivo* gene expression profiles of uro-pathogenic *E. coli* isolates. Furthermore, to construct a phylogenetic tree based on the core genes and to assign the UTI-associated *E. coli* isolates to different phylogenetic groups; to find the global gene expression pattern by using the pan-genome based transcriptomic pipeline. Similarly, to gain insights on the variation of *K. pneumoniae* transcriptional landscape, the same pipeline is used.

As a second aim, I generate the consensus nucleotide sequence of the *P. aeruginosa* core genome in order to identify SNPs at high accuracy and to find the patho-adaptive mutations in *P. aeruginosa* clinical isolates. Further aims are, i) to exclusively classify the clonal-related single nucleotide polymorphisms (SNPs); single SNPs those are only present in one isolate, and multi-lineage specific SNPs (inter-clonal) and ii) additional classification of synonymous and non-synonymous mutations of inter-clonal SNPs to define the patho-adaptive mutations of clinical *P. aeruginosa* isolates. Tracking these recurrent patterns of clonally independent mutations is increasingly important to understand the adaptation strategies of *P. aeruginosa*, especially in the virulence niches. The genomic information of clinical *P. aeruginosa* isolates will be captured by applying whole genome sequencing. The consensus nucleotide sequence based SNP identification will be applied to describe the genetic variations at single nucleotide level to understand the adaptation strategies and evolutionary processes of the clinical *P. aeruginosa* isolates.

Results

In vivo mRNA profiling of uropathogenic *Escherichia coli* isolates

1 In *vivo* mRNA profiling of uropathogenic *Escherichia coli* isolates

During my thesis, the study "*In vivo* mRNA profiling of uropathogenic *Escherichia coli* from diverse phylogroups reveals common and group specific gene expression profiles" was published in mBio. Piotr Bielecki and I are the co-first authors in this publication. This publication analyzed transcriptomic data of 21 uropathogenic *E. coli* (UPEC) strains by using pan-genome of *E. coli* as basis, to acquire information on the expression of *E. coli* pathogenicity genes during urinary tract infections (UTI) in humans.

Strand-specific RNA-sequencing was used to generate comprehensive *in vivo* transcriptional profiles of 21 UPEC strains causing symptomatic UTI in a cohort of elderly patients. Since *E. coli* strains are divided into several phylogroups [38, 39], it is important to use multiple reference strains of *E. coli* in order to map all reads from the diverse clinical UPEC strains. Hence, the major objective of my work in this study was i) to include all the publicly available strains of *E. coli* to create one reference genome (pan-genome) for mapping a maximum number of reads, ii) to construct a phylogenetic tree based on the core genes in order to assign the UTI-associated *E. coli* isolates to different phylogenetic groups, and iii) to identify the global patterns of gene expression of all strains. Figure 8 depicts a pan-genome based transcriptomic data analysis pipeline.

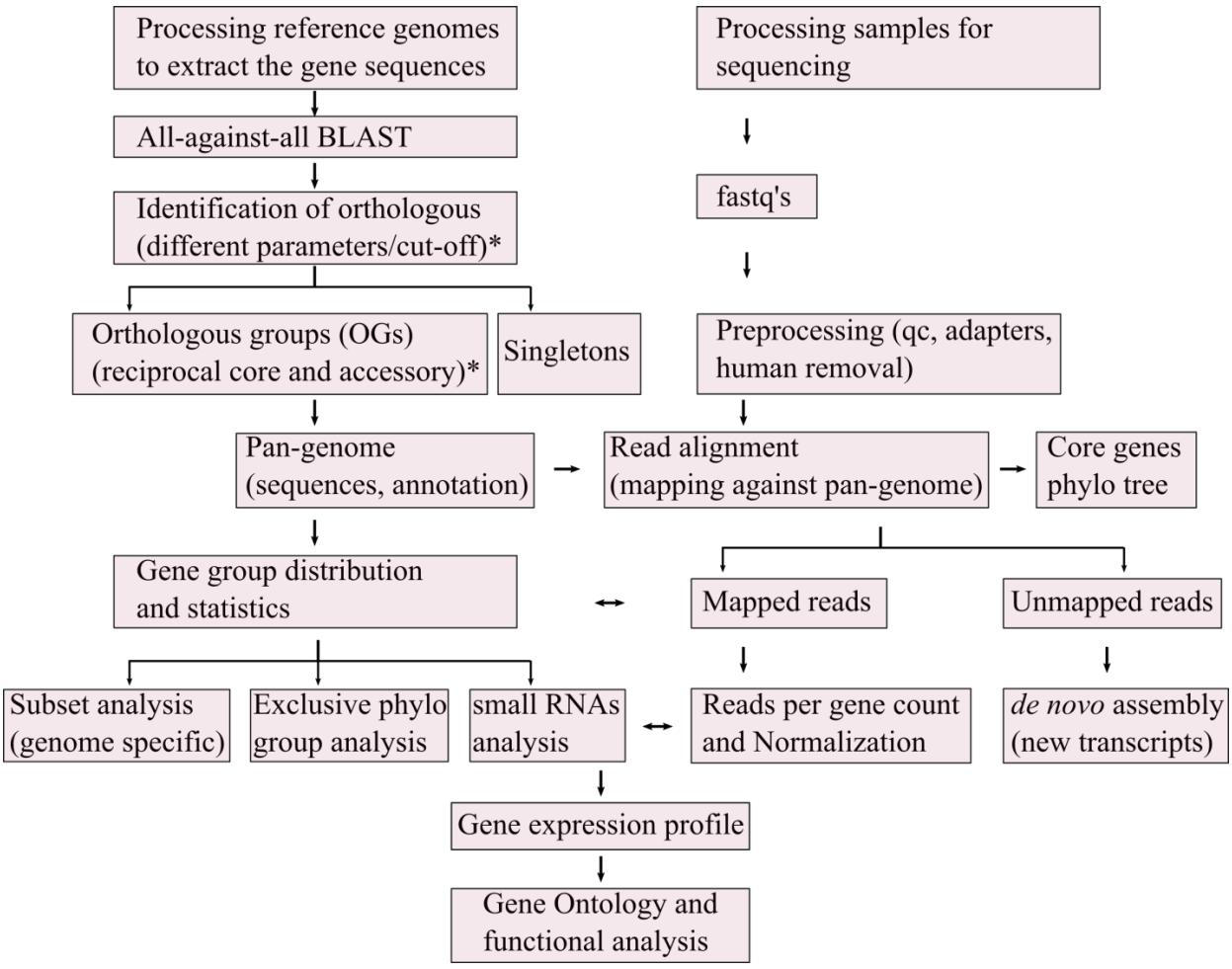


Figure 8: Flow diagram of pan-genome based RNA-seq data analysis pipeline.

Different cut-off values can be applied to evaluate the core genome as well as the total number of non-redundant gene families, see (star *) in the diagram. Scripts are written in Perl; R and DESeq are used in differential gene expression analysis.

I downloaded and used the 54 strains of *E. coli* (more details in table 1) that were publicly available from Genbank/EMBL to build the pan-genome.

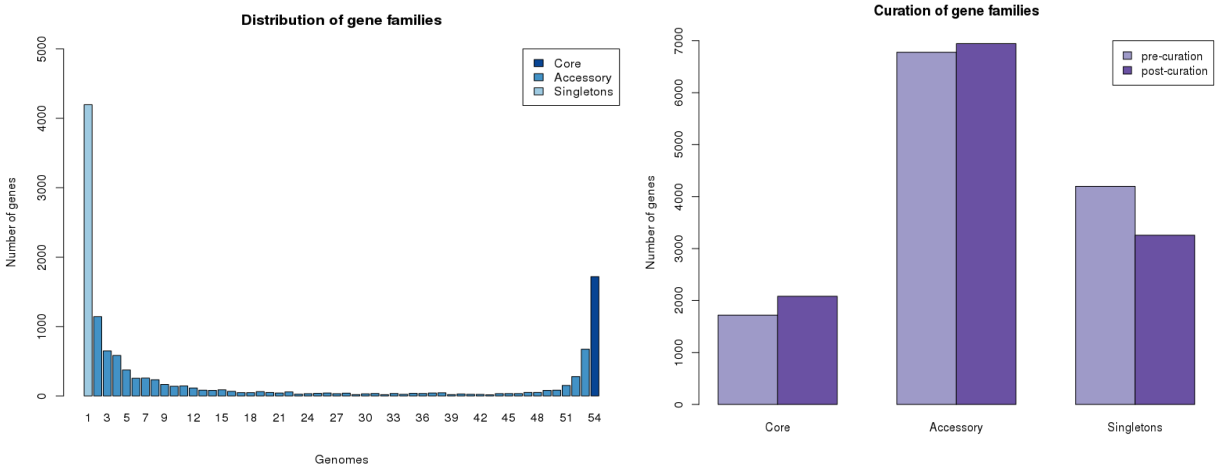
Table 1: List of fully sequenced genomes in *E. coli* as on September 30, 2011

Accession id	***E. coli* Genome name**	**Number of genes**
CP000468	APEC O1	4430
AM946981	BL21 (DE3)	4204
U00096	K-12 substr. MG1655	4146
CP001396	BW2952	4084
CP001925	Xuzhou21	5039
CP003034	O7:K1 CE10	5009
AE014075	CFT 073	5369
FM180568	O127:H6 str. E2348/69	4552
AP010958	O103:H2 str. 12009	5054
CP003109	O55:H7 str. RM12579	4912
CP000247	str. 536	4619
AP009378	str. SE15	4338
BA000007	O157:H7 str. Sakai	5230
CP002516	KO11FL	4533
CP002211	clone D i2	4919
CU651637	LF82	4376
CP002167	UM146	4650
CP002967	str. W	4608
FN554766	str. 042	4800
CU928145	str. 55989	4759
CP001671	ABU 83972	4793
CP001665	BL21-Gold (DE3) pLysS AG	4228
CP000819	REL606	4204
CP000948	K-12 substr. DH10B	4128
AP012030	DH1 (ME8569)	4259
CP001637	str. DH1	4160

AP010960	O111:H- str. 11128	4972
CP001969	IHE3034	4757
CU928161	str. S88	4692
CP000970	SMS-3-5	4742
CU928163	UMN026	4823
FN649414	ETEC H10407	4697
CP002212	clone D i14	4919
CP001855	O83:H1 str. NRG 857C	4429
CP002729	UMNK88	5117
AP009048	K-12 substr. W3110	4217
CP001509	BL21 (DE3)	4157
CP000800	E24377A	4749
CU928162	ED1a	4915
CP001164	O157:H7 str. EC4115	5315
CP000802	str. HS	4377
CU928160	str. IAI1	4349
CU928164	str. IAI39	4730
CP002797	str. NA114	4873
AP010953	O26:H11 str. 11368	5364
CP000946	ATCC 8739	4199
AP009240	str. SE11	4675
CP001368	O157:H7 str. TW14359	5255
CP002185	str. W	4478
CP001846	O55:H7 str. CB9615	5014
CP002970	KO11FL	4697
CP002291	P12b	4393
CP000243	UTI89	5017
AE005174	O157:H7 str. EDL933	5298
Total number of genes		252623

The 54 *E. coli* strains contain 252,623 genes, which give an average of 4,678 genes per strain. With the aim to collapse those genes into gene families and define the genes present in all genomes (core genome), I first extracted all coding sequences (CDS) from the corresponding genomes, then blasted all-against-all using BLASTP [10], selecting hits with greater than 90% length and 50% sequence identity. Only if a gene product had a maximal reciprocal set of homologs in all other strains, 54 in total, the corresponding gene was considered "core"; otherwise, it was considered "flexible". I re-evaluated the flexible CDS that had homologs in 53 or 52 of the 54 *E. coli* genomes. The set of core genes detected in the reciprocal blast search comprised 1,719 CDS, while there were an additional 363 CDS manually assigned to the core genome, summing to 2,082 core CDS (see appendix for more information). Apart from the 2,082 core CDS; I identified 10,202 flexible CDS, including 3,257 singletons.

I also extracted the genomic sequences of 70 noncoding RNAs (ncRNAs) from the O26:H11 genome and performed BLASTN searches against each of the 54 genomes in order to define how many of these ncRNAs are present in all *E. coli* genomes. A total of 47 ncRNAs were found in all 54 genomes and these ncRNAs were included in the core genome that consisted of 2,129 (2,082 CDS and 47 ncRNAs) genes. Finally, the sum of core (2,129) and flexible (10,202) genes amounted to 12,331 genes. This initial distribution and a further curation of these gene families are shown in Figure 9.



(a) (b)

Figure 9: Distribution and curation of gene families.

(a) The distribution of gene families shared in any number of genomes of *E. coli* (1 to 54) is presented; number of core genes is highlighted in dark blue, accessory genes in blue and singletons in light blue. (b) Before and after curation of gene families in core-, accessory-genomes and singletons.

The raw Illumina sequence reads (36-bp single end) were first split according to their bar codes using the *fastq-mcf* script of the ea-utils package (https://code.google.com/p/ea-utils/), and then adapter and bar code sequences as well as low quality sequences were removed. *bowtie-build* module in the Bowtie package were used to build an indexed reference based on the 12,331 *E. coli* genes found in the 54 reference genomes. Mapping to the pan-genome was performed using Bowtie with options "-m 1 -best -strata" to allow only uniquely mapping hits and avoid uncertainties regarding repeat regions and ribosomal genes. Read counts were extracted from the SAM output files for each annotated gene and were used as an input for differential gene expression calculations with the R package *DESeq* [96]. The read counts per gene (RPG) were normalized per kilobase of gene sequence as described by Dötsch *et al*. [97] values according to the following equation:

$$\text{nRPK} = log2\left(\frac{1000}{li} * \frac{RPG}{Fj} + 1\right)$$

where *li* is the length in bp of gene *i*, *RPG* is the absolute count of reads of gene *i* and *Fj* is the size factor calculated by DESeq of isolate *j*. This normalization method delivers more robust data as e.g. RPKM when analyzing highly expressed genes [97]. Genes were considered to be differentially regulated only if their absolute logarithmic fold change over the control was higher than 1 at a false discovery rate of a maximum 5% (Benjamini and Hochberg *P* value correction provided in *DESeq*). Since the list of genes (pan-genome that includes core-, accessory- genes and singletons) is quite big (also with orthologous identifiers and/or expression values in nRPK, normalized reads per kilobase) as used in the different analysis, they are presented online as supplemental Data Sets (see appendix).

Mapping of all obtained reads to the pan-genome revealed that only very few core genes were not transcribed in any of the 21 UPEC isolates, indicating that expression of most of the core genome is relevant for bacterial replication in the human urinary tract. Further, a large set of overall 2,589 genes were found that were commonly expressed in all isolates (accounts for 52% to 67% of all transcribed genes within one isolate). They appear to be unregulated or constitutively expressed as the overall variation of the expression profiles among the isolates was low and the genes were generally expressed at a high level independently of their phylogenetic group specificity (Figure 10). As expected, many of these genes correspond to genes required for the maintenance of basic cellular functions, such as DNA repair, ATP synthesis, amino sugar metabolism, and protein transport (see appendix).

Mapping of all obtained reads to the pan-genome revealed that – apart from the commonly transcribed genes (2,589 genes), a large fraction of 6,305 genes were expressed in at least one of the 21 UTI samples. The expression profile of the 21 UTI samples clustered into three main groups that represented the B2, D, and A/B1 phylogenetic groups (for phylogenetic tree, see Fig. 1 in the publication). Of note, clustering became even more accurate and well separated when only the expression of genes of the accessory genome was included in the analysis (Figure 11). These results are in agreement with previous reports [39, 40] and clearly demonstrate that the presence of group specific gene repertoires (and not a difference in overall gene expression profiles) impacts on clustering of the UTI isolates into the phylogroups.

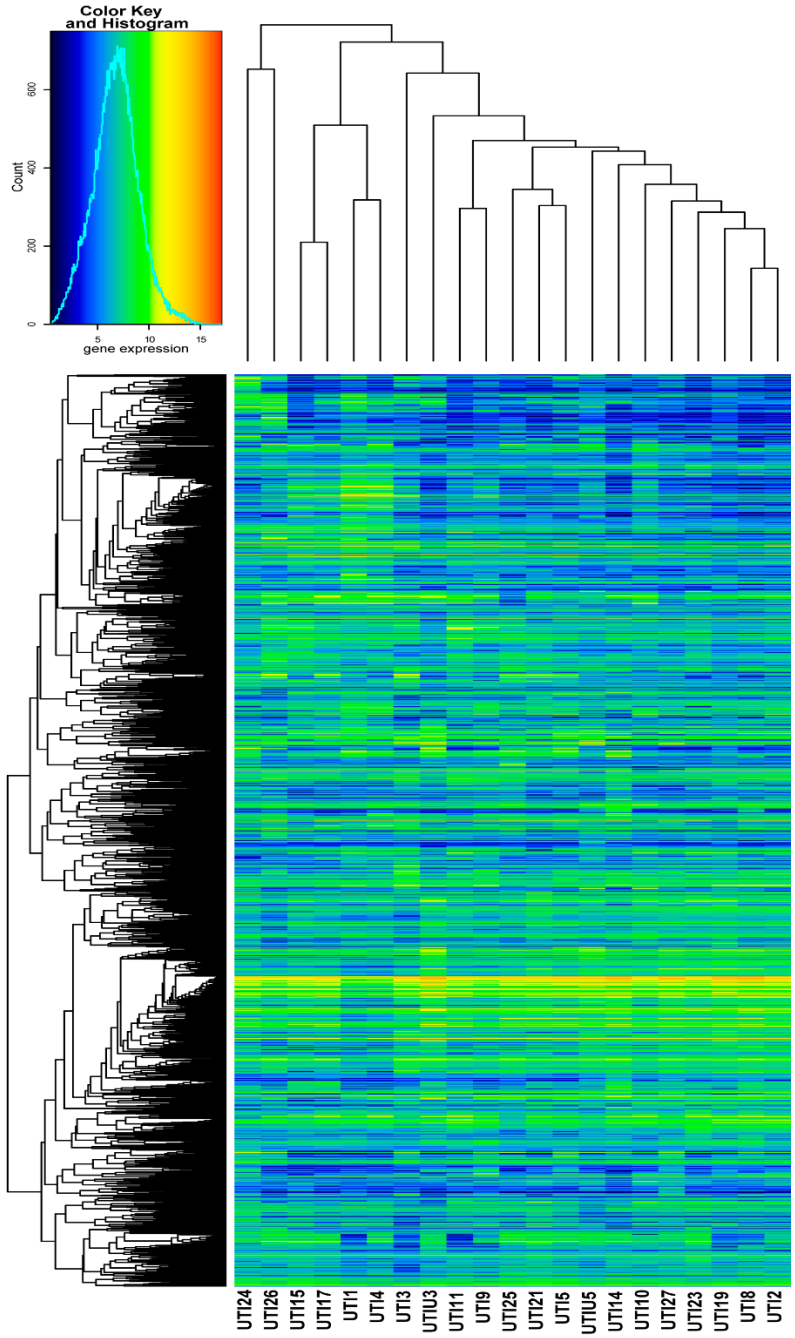


Figure 10: Expression of the 2,589 commonly transcribed genes within the 21 clinical UPEC isolates

The heat map show the expression of the (2,589) commonly transcribed *E. coli* genes within the 21 clinical isolates (in nRPK). The genes (vertical) are hierarchically clustered using Pearson distances, and the isolates (horizontal) are clustered according to Spearman rank correlation. The histogram describes the correlation of the color to the nRPK value of absolute expression.

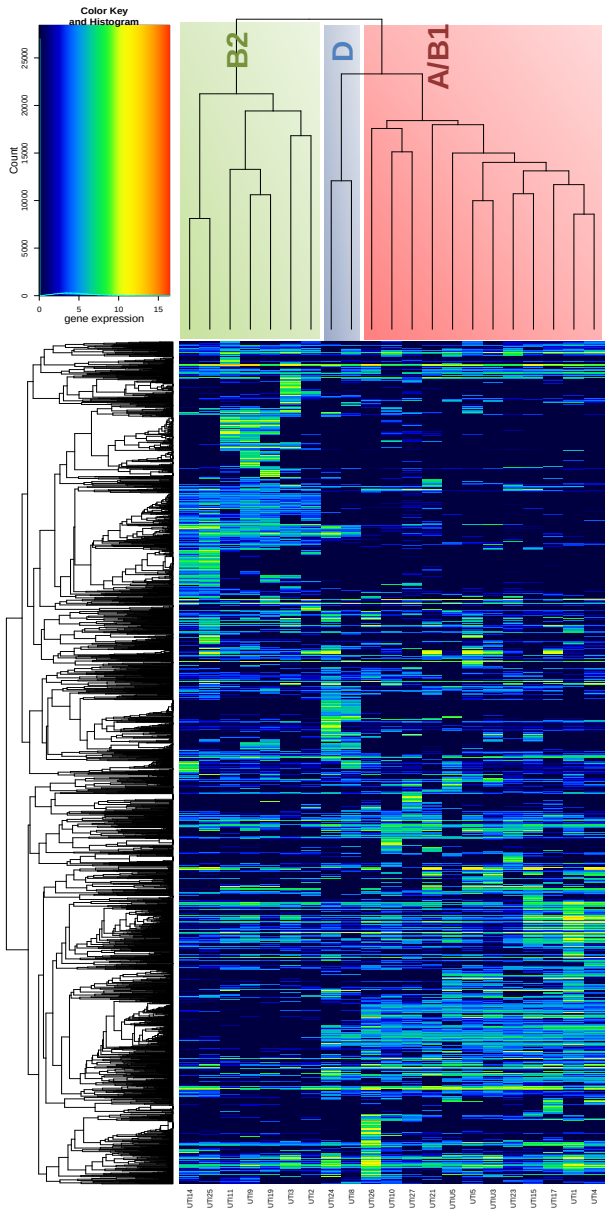


Figure 11: Expression of the *E. coli* accessory genome among 21 clinical UPEC isolates.
The expression of genes of the accessory genome correlates to the phylogenetic groups. The genes (vertical) are hierarchically clustered using Pearson distances, and the isolates (horizontal) are clustered according to Spearman rank correlation (nRPK values used as above). Only genes that show a variance greater than 2 between the isolates are included (2,409 genes).

Many studies have demonstrated that the phylogroups differ with respect to the presence of virulence factors and colonization of ecological niches, and UPEC isolates have previously been found to be more prevalent in phylogroup B2 [41]. 7 UPEC isolates that grouped with B2

phylogenetic group were found, and they expressed several virulence genes *in vivo* that have been associated with UPEC strains exhibiting full-pathogenic potential. This also included a novel set of genes overrepresented in those isolates (see appendix in publication). Nevertheless – and in accordance with previous studies on atypical UTI patient populations [42-44] – in our study, which was performed on samples collected mainly from elderly patients, as many of 12 out of the 21 UPEC isolates analyzed were assigned to the A and B1 phylogenetic groups, which predominate among commensal *E. coli*. Here, I identified 142 genes that were expressed at a significantly higher level in the 12 phylogroup A/B1, isolates compared to all other 9 isolates (Figure 12).

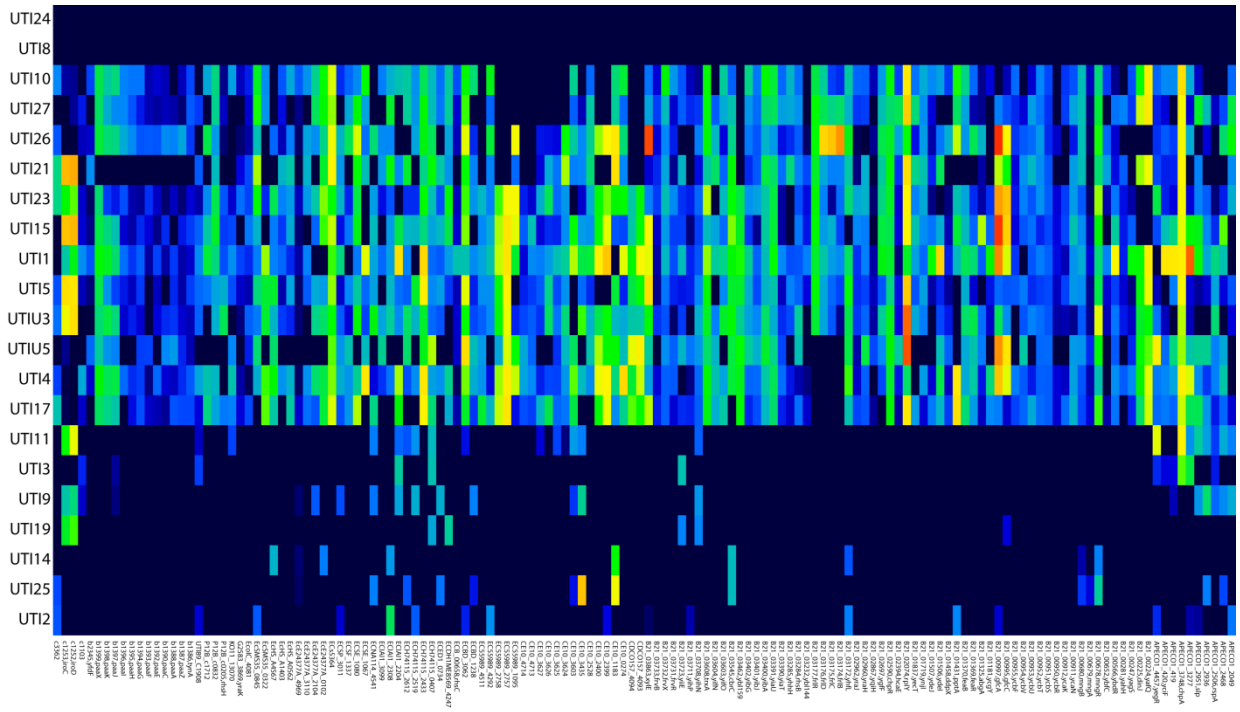


Figure 12: Expression of phylogenetic group A/B1 specific genes among 21 UPEC isolates. The heat map show only those genes that are expressed in >70% of the A/B1 phylogroup-specific isolates and in not more than 30% of the isolates from other phylogroups within 21 clinical isolates are included.

In conclusion, I found that although *E. coli* isolates that have been assigned to the four phylogenetic groups (A, B1, B2 and D) share a large general gene expression profile, they do express clearly distinct accessory genomes. A strong correlation between the *E. coli in vivo* expression of the accessory genome and the genetic background of the isolate were found. However, as has been described before [39, 40], this correlation was dependent on the acquisition of group specific genes in the accessory genomes rather than on a difference in their expression profile, possibly reflecting their evolution in distinct niches. Furthermore, a novel set of genes were identified that were exclusively expressed in the 7 UPEC isolates clustering with group B2; and in addition, a set of 142 genes whose expression was demonstrated to be specifically enriched in the 12 isolates that clustered with the A/B1 phylogroups. Hence, the pan-genomic approach allowed the identification of the gene functions that might be responsible for an acquired phenotype and, for the variation landscape of the core genome in order to identify the impact of genetic variation on the phenotype.

The following article has been published in mBio. Piotr Bielecki and myself are co-first authors. The DOI of the publication is 10.1128/mBio.01075-14 and it is available on line at: http://mbio.asm.org/content/5/4/e01075-14.full [88].

Transcriptome analysis of clinical *Klebsiella pneumoniae* isolates

2 Transcriptome analysis of clinical *Klebsiella pneumoniae* isolates

The pan-genome approach has become a standard practice to study the population structure of bacterial species. Further, the pan-genome serves as a reference for read mapping, qualitatively and quantitatively (as described in the previous study). To determine the conservation and variation of gene expression profiles across 37 clinical *K. pneumoniae* isolates, the pan-genome of *K. pneumoniae* is a valuable tool. The study "Deep transcriptome profiling of clinical *Klebsiella pneumoniae* isolates reveals strain and sequence type-specific adaptation" was published in Environmental Microbiology. Sebastian Bruchmann is the first author of this study, I am the second author, involved in the development and analyses of the *K. pneumoniae* pan-genome, the prediction of small RNAs and read mapping.

RNA-sequencing was used to generate the transcriptomic profiles of 37 clinical *K. pneumoniae* isolates of various phylogenetic origins [47]. The main objective of this research was to gain insights on the variation of *K.* pneumoniae transcriptional landscape by using the established pan-genome based transcriptomic analysis pipeline.

Similarly as described in the last chapter, I developed the pan-genome of *K. pneumoniae* based on the 11 publicly available *K. pneumoniae* genomes (more information on table 2). *K. pneumoniae* shows a high variation in genome content like many other gram-negative bacteria [45]. The 11 reference genomes contain 57,312 genes which vary from 5.2 to 6.1 Mbp, encoding for 4,887 to 5,577 genes per genome. I first extracted all the gene sequences from the corresponding genomes and then blasted them against each other (all-against-all) using BLASTN, selecting hits with greater than 90% sequence length and 90% sequence identity. This pan-genome contained 7,859 genes of which 3,336 genes were shared by all genomes (core genome) and 4,523 genes were absent in at least one of the genomes. Among the latter, 1,598 genes were identified in only one of the reference genome (singletons). Further, I attempted to predict the small RNAs in *K. pneumoniae* subsp. MGH78578 (in addition to 51 previously annotated sRNAs) by using sRNAscanner [46] for small RNAs expression analysis. A total of 30 small RNAs were predicted with the default parameters and included in the total non-redundant gene repertoire (7,859 genes).

Table 2: List of fully sequenced genomes in *K. pneumoniae* as on April 30, 2014

Accession ID	K. pneumoniae strain	Number of genes
AP006725	NTUH-K2044	5123
CP000647	MGH 78578	4887
CP002910	KCTC 2242	5035
CP003200	HS11286	5404
CP003785	1084	5067
CP003999	Kp13	5299
CP006648	CG43	4897
CP006656	JM45	4980
CP006918	30684/NJST258_2	5545
CP006923	30660/NJST258_1	5577
CP006659	ATCC BAA-2146	5498
Predicted small RNAs		30
Total genes (including 30 predicted small RNAs)		57342
Pan-genome		7859
Unique genes (singletons)		1598
Core-genome		3336
Accessory genes		2925

I also created an orthologous matrix based on the information of presence and absence of genes in every genome, for statistical extrapolation analysis. Figure 13 depicts the development of the size of the pan-genome, core-genome and singletons by sequentially adding the genomic information of one genome to that of the others. Based on extrapolated data, it is expected that the size of these groups would change by less than 2% if genomic information of another genome would be added, less than 1% by adding information of 20 genomes, and less than 0.5% by adding 40 genomes in total.

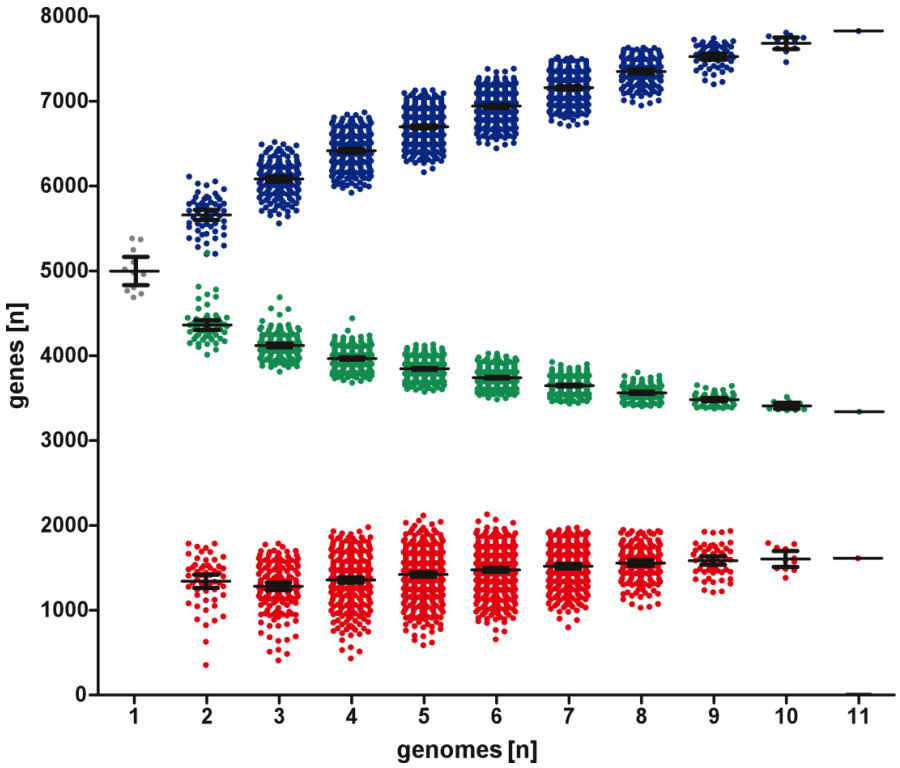


Figure 13: Analysis of *Klebsiella pneumoniae* genomic content.

The amount of genes belonging to the pan-genome (blue dots), core-genome (green dots) and unique genes (red dots) is plotted as a function of genomes sequentially added in all possible combinations. The number of genes from the first genome is shown as grey dots. Black bars show mean with 95% confidence intervals.

The core genome sequences were extracted from the strain type MGH78578. To exclude any bias in mapping efficiency due to the choice of the core genome strain type MGH7858, I also created a pan-genome based on a different sequence type ST38. The comparison of mapping efficiency revealed an average difference of just 2000 reads (or 0.001%) when using the two alternative pan-genomes. Up to 99% of the reads could be mapped to the non-redundant gene repertoire of the *K. pneumoniae* pan-genome. Of all 7,859 genes in the pan-genome, a large set of 3,346 genes were identified that were commonly expressed in all isolates above the threshold level and normalization was done as described in the last chapter. This core transcriptome accounted for 62% to 71% of all transcribed genes within one isolate and largely overlapped with the core genome (75% of the core genome was commonly transcribed). Due to this large

overlap between core genome and core transcriptome, it was not surprising that the core transcriptome likewise consisted mostly of genes with housekeeping functions.

I next assessed whether and how the genetic background of the various clinical isolates impacts on global gene expression profiles. I performed a hierarchical clustering based on the Spearman rank correlation of all the isolates according to the overall similarity of the expression of all 7,859 genes in the pan-genome. Hierarchical clustering revealed three major subgroups: the first included the two ST512 isolates together with all the ten ST258 isolates; the second group included all except one ST101 isolates, and the third group included the four ST15 isolates together with various other sequence types. This clustering may have been expected since the genomic composition, i.e. the presence of distinct sets of accessory genes of the various MLST sequence types strongly influences the clustering. To evaluate the transcription of accessory genes, the clustering of the 37 clinical isolates based on variations within genes of the core-transcriptome were analyzed. Remarkably, clustering of expression profiles based solely on the core-transcriptome still revealed concordance with the phylogenetic groups (see Fig. 4 in publication).

In conclusion, this study revealed that the core transcriptome of the *K. pneumoniae* isolates (comprises 3,346 genes) and that global core gene expression profiles are similar among but distinct between the various sequence types (STs). Furthermore, our analysis of how differentially expressed genes correlate to clinically relevant phenotypes such as biofilm formation and virulence uncovered that *K. pneumoniae* sequence type-specific traits might help successful spreading or survival of the epidemic clonal lineage in the hospital setting.

The following article has been published in Environmental Microbiology. Sebastian Bruchmann is the first author and I am the second author in this study. The DOI of the publication is 10.1111/1462-2920.13016 and it is available on line at: http://onlinelibrary.wiley.com/doi/10.1111/1462-2920.13016/full [89].

Genome-scale analysis of genetic diversity in *Pseudomonas aeruginosa* populations – an orthologous group based consensus approach

3.1 Introduction

A major challenge in biological and environmental research is to understand the intricate link between genotypes and phenotypes and how the environment influences that link. Genetic variants such as mutations or the presence and absence of a gene among individuals of a population can produce different phenotypes that are selected for or against by evolution [48]. Thus, evolutionary change critically relies on genetic variations. The genetic variation of an entire species is often called genetic diversity. A genetically diverse population is usually characterized by more phenotypic variation. This can be advantageous because it enables some individuals and, therefore, a population, to survive in changing environments.

A promising approach to understand the evolution of genetic diversity is to study genetic changes within populations that have to adapt to novel and challenging habitats. Those habitats are sub-optimal for sustained growth and thus provide strong selection for adaptive changes. *Pseudomonas aeruginosa* represents an excellent model for understanding the molecular mechanisms of the adaptation of a bacterial species to very challenging habitats. The opportunistic human pathogen inhabits a stable source-habitat in the natural environment, which serves as a *reservoir* to occasionally cause infectious diseases in humans. If not eradicated during the acute infection phase, *P. aeruginosa* populations can cause chronic infections and thus exist for a long enough time for evolution to take place [49].

In order to learn more about evolutionary processes, the identification of the genetic determinants that underlie the selection of favorable phenotypes is critical. The precise determination of genetic variants mostly begins by mapping DNA sequencing reads to a reference, which is used to represent the species. This can be either a reference genome or a population consensus read alignment. The reference acts as a common coordinate system for the majority of the genome and the mapping approach is an effective way of detecting genetic variation caused by single nucleotide polymorphisms (SNPs) between a sample and the reference [50, 51]. However, this approach critically relies of the availability of a high quality reference. Furthermore, it is less effective when the analyzed genomes considerably differ from the reference. The sample genome may contain flexible sequences, such as horizontally acquired

genes, that are not represented in the reference and the sequences might substantially diverge from the reference, so that full identification of genetic variation becomes a challenge.

Due to this reference allele bias, studies on e.g. the within-host evolution of *P. aeruginosa*, which persists for years in the lung of cystic fibrosis patients, have been restricted to the analysis of clonal lineages: adaptive patterns of sequence variation in the genes targeted by natural selection could be uncovered, because the genome of the early infecting *P. aeruginosa* isolate served as a reference for the subsequently recovered *P. aeruginosa* strains [52-57].

Here, I aimed at addressing the problem of the reference allele bias and explored whether pan-genomic information could better define the full genetic variation within a group of phylogenetically diverse *P. aeruginosa* isolates. I therefore explored creating a pan-genome core gene reference, which was based on annotated genes across a set of 99 genotypic diverse clinical *P. aeruginosa* isolates. Indeed, the development of a rich reference structure allowed robust detection of the full genetic diversity within the 3,814 core genes of our clinical *P. aeruginosa* strain collection. Furthermore, taking phylogenetic information into account, I was able to define sequence variations that were associated with certain phylo-groups and others that evolved independent on the phylogenetic background of the clinical isolates. Our results clearly demonstrate the usefulness of the pan-genome approach to describe genetic diversity within a bacterial species. Applying this approach to the strict core genes of a collection of 99 clinical isolates furthermore revealed that the *P. aeruginosa* core genome is highly conserved across clinical isolates and in general not subject to adaptive evolution.

3.2 Results

3.2.1 The *Pseudomonas aeruginosa* pan-genome

Whole genome sequencing of a multitude of bacterial isolates from one species gives detailed information on the available gene pool and includes information on the number of genes that is common to all of the isolates of one species (core genome) and the number of genes that can be found in only sub-fraction of isolates (accessory genome) or even only in single strains (singletons) [58]. In this study, the genomes of a total of 99 clinical *P. aeruginosa* isolates were sequenced by the use of Illumina technology. Those clinical isolates have been sampled from various infection sites from patients at different hospitals across Germany. I performed *de novo* assembly of the genomes and included the genomic information of two *P. aeruginosa* reference strains, PA14 and PAO1, for the generation of the pan-genome. *De novo* annotation revealed an overall gene content of 6,255 genes on average among the *P. aeruginosa* strains included in the analysis with a distribution from 5,677 to 6,831 genes per genome. To define the core genes, I collapsed these genes into orthologous groups by blasting all-against-all gene sequences using BLASTN [10]. Only if a gene had a maximal reciprocal set of orthologous in all other strains, 101 in total, the corresponding gene was considered core; otherwise, it was considered accessory. The set of orthologous genes - from strict core to singletons - yielded a cumulated non-redundant gene number of 18,319. While 14,505 genes belonged to the accessory genome (5,539 of which were singletons), 3,814 were core. Of note, within the group of accessory genes a large number of genes (1,257) were present in almost all genomes, in other words, they were absent in only 1-5% of the total isolates (supplementary figure S1). Closer manual inspection revealed that many of those genes were not completely absent in the 1-5% of the isolates. Instead, they were affected by partial gene losses, indicating that those genes are core genes, which acquired secondary inactivating mutations. Thus, I consider the 1,257 genes as soft core genes. Of note, among those soft core genes there might be some truly accessory genes, because they are completely absent in a small fraction of isolates. However, there might also be additional soft core genes. For example, genes such as *lasR* show secondary (incomplete) gene losses in more than 5 % of our clinical isolates and therefore are not considered as soft core genes in our analysis. Thus, the number of 1,257 soft core genes can only be an estimate.

The main features of the *P. aeruginosa* pan-genome of our 99 isolates were comparable to those of previously published pan-genome analyses. The high numbers of strict (3,814) and soft (1,257) core genes corroborates the previous finding that the *P. aeruginosa* genome is very conserved [59-62]. A graphical representation of the correlation of the number of sequenced genomes and the increase in the size of the pan-, and singleton-genome is shown in Figure 14. The saturation model estimates that with the information of 685 genomes 95% saturation could be reached, which would correspond to a pan-genome size of 25,882 genes including 3,062 strict core genes and 15,765 accessory genes, of which 7,055 would be singletons (supplementary figure S2).

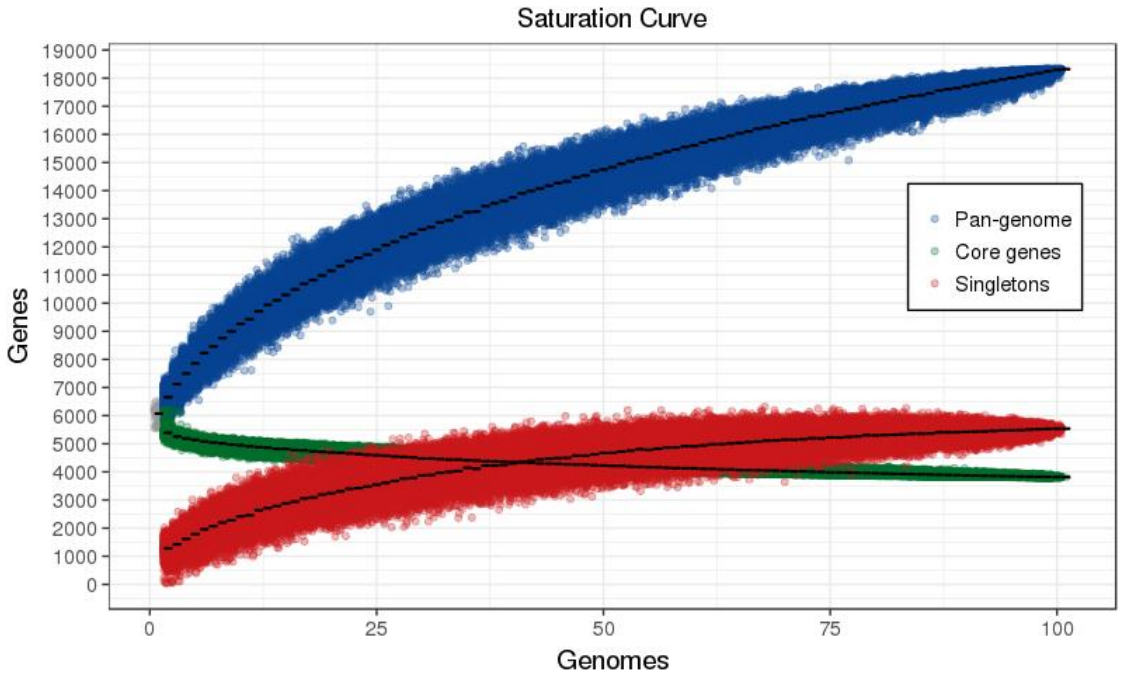


Figure 14: Influence of the number of sequenced genomes on the *P. aeruginosa* pan- and core-genome size.

99 clinical *P. aeruginosa* isolates and 2 reference strains are used (n=101). The total number of genes belonging to the pan-genome (blue dots), core genome (green dots) and singletons (red dots) is plotted as a function of genomes sequentially added in all possible combinations. Based on the averages, it shows an exponential expansion for all three groups. Genes from the first genome is shown in grey dots. Black bars show mean with 95% confidence intervals.

3.2.2 Phylogenetic relationship of the clinical *P. aeruginosa* isolates

We constructed the phylogenetic tree of the 99 clinical isolates with a neighbor-joining algorithm based on a distance matrix calculated from *k*-mers of the 3,814 core genes. The core genome which includes the seven standard multi locus sequence type (MLST) genes of *P. aeruginosa* (*aroE, trpE, guaA, nuoD, ppsA, acsA* and *mutL*) [63] represents more than 61% of the average *P. aeruginosa* genome size, and thus allows for a very fine-scale resolution of clonal linages. This phylogenetic tree, in agreement with previous studies, largely consists of two major non-overlapping clusters, which contain the PA14 and the PAO1 type strains respectively [64, 65] (Figure 15). In our strain collection, 44 clinical *P. aeruginosa* isolates clustered with the PA14 phylogroup and 55 isolates with the PAO1 phylogroup. The tree also provides information on more than 40 distinct sequence types (STs) that includes major sequence types ST235, ST111 and ST132. These subgroups were identified based on the MLST profiles of *P. aeruginosa*. The phylogenetic distribution of our 99 clinical isolates was found to be comparable to the phylogenetic diversity of 52 previously fully sequenced *P. aeruginosa* strains whose sequence information is publicly available in GenBank/EMBL (supplementary figure S3).

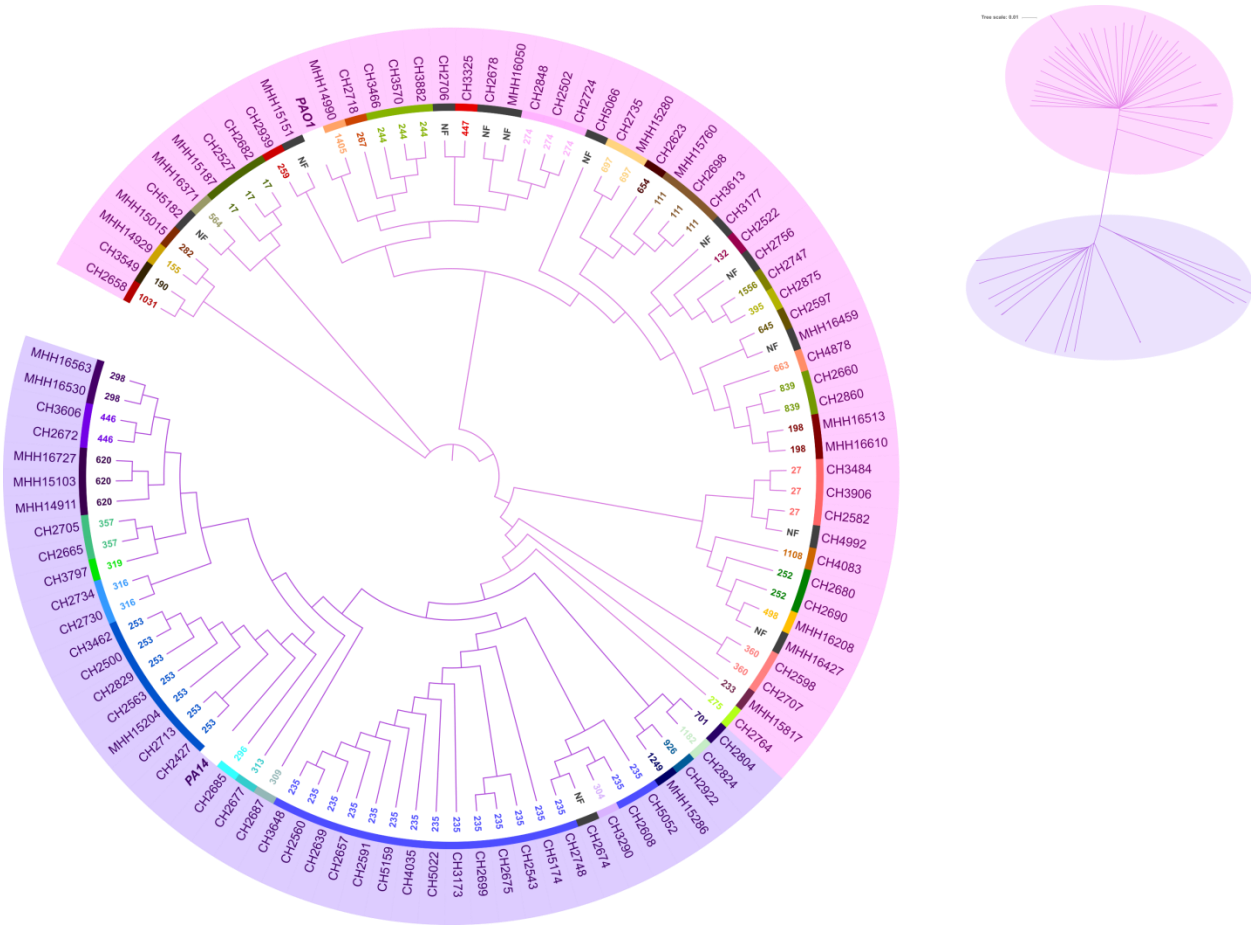


Figure 15: Broad phylogenetic distribution of the 99 *P.* aeruginosa clinical isolates.

The unrooted tree on the top right indicates that the isolates separate into two clonal complexes (PA14-like and PAO1-like). The circular tree highlights in pink the PAO1-like isolates and in violet the PA14-like isolates. Sequence type (ST) information is shown as an inner ring label. NF indicates newly identified STs that are not found in the database.

3.2.3 Consensus sequence of the *P. aeruginosa* core genes

Instead of mapping sequencing reads to a reference, I explored the full genetic diversity of *P. aeruginosa* core genes by establishing a gene-wide consensus sequence, which gives information on the most frequent nucleotide at every single position across the homologous genes that have diverged in different isolates from a common ancestral gene (orthologous groups) (Figure 16). I concentrated on those orthologous groups for which sequence information of all 99 isolates was available (core genes orthologous groups). Among the 3,814 core genes, 3,014 genes were identical in length within their orthologous groups. For the 800 genes with non-identical length, standard multiple sequence alignment was performed to identify the gaps, that were then filled with an "N" to have an identical number of nucleotide positions across all isolates. Those orthologous groups then served as the basis for the generation of a consensus sequence of the respective core gene in which information on the most frequent nucleotide at all positions across all isolates is calculated. Once the consensus sequence was generated, SNPs were scanned for each position across all the isolates. An example of the consensus sequence is shown in supplementary figure S4. The visualization of the genetic diversity at the single nucleotide level of *P. aeruginosa* core genes together with the consensus sequence is provided in the Bactome database (accessible through https://bactome.helmholtz-hzi.de).

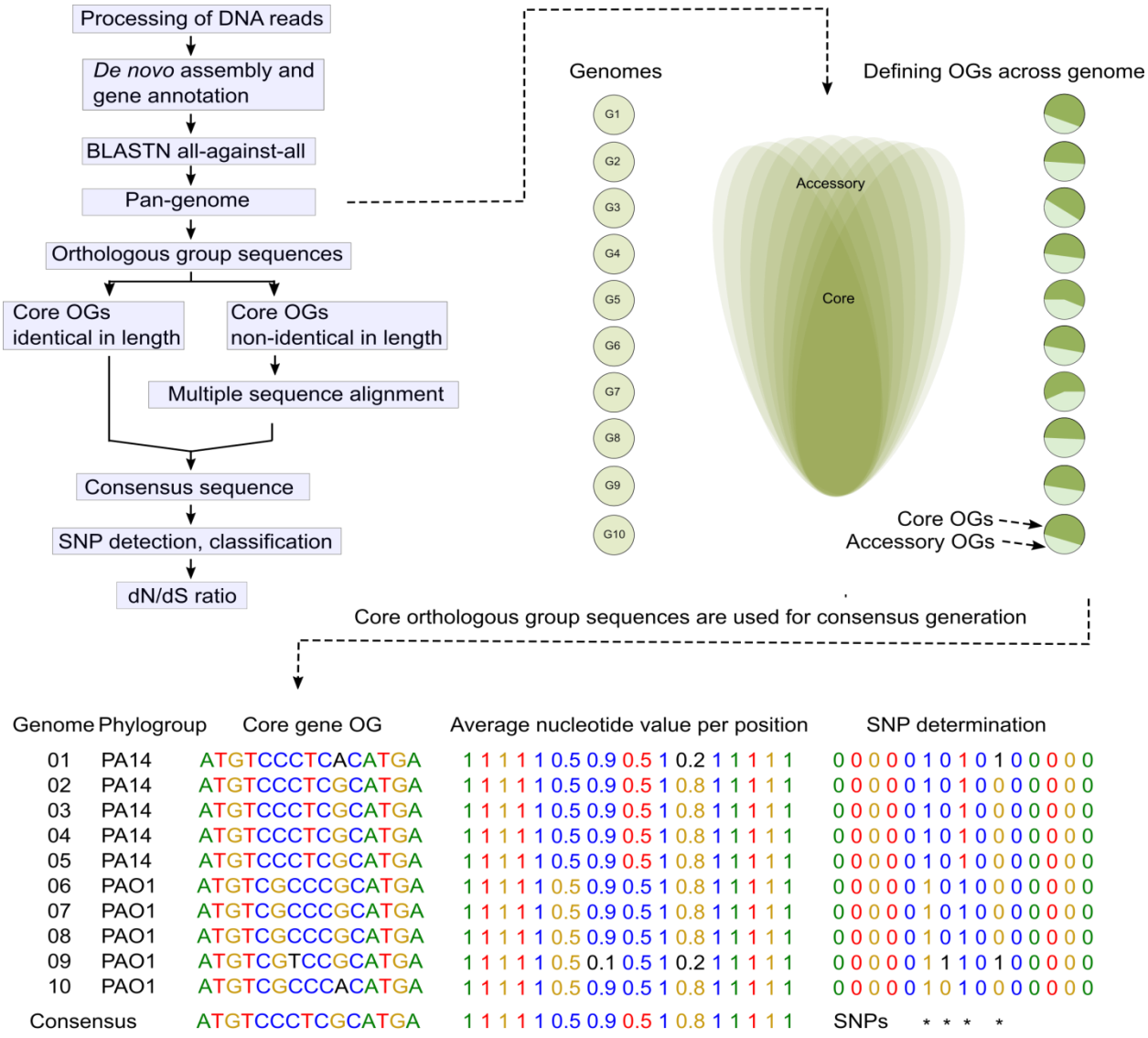


Figure 16: Framework for the gene-wide consensus sequence generation and SNP classification.

Raw DNA sequencing reads are processed and error corrected before *de novo* assembly and gene annotation. Orthologous groups (OGs) are identified based on an all-against-all sequence comparison of the genes in the assembled genomes. Each core gene orthologous group is aligned to create the consensus nucleotide sequence across the genomes along with the average value of the nucleotides. Any individual position that deviates from the consensus is determined as a SNP. Phylogenetic information is further used to classify the clonal-specific, single- and inter-clonal SNPs.

3.2.4 Identification of SNPs in the *P. aeruginosa* phylogroups

By using the pan-genome approach, SNPs within the 3,814 core genes could be identified in our *P. aeruginosa* population without the need for a reference to which sequences are mapped. A total of 159,609 SNPs were identified across all the isolates in 3,814 genes spanning 3,629,979 positions in every genome. This corresponds to sequence diversity at the single nucleotide polymorphism level of 0.04 in the core genome, indicating that those genes are highly conserved and differ by just a few SNPs from one isolate to the other [64, 66, 80]. Only, seven genes – namely, *rpsS, rpmC, minE, acpP, lppL,* PA14_07370 and PA14_12560 – were completely conserved and did not harbor any SNPs across the isolates. I found 49,722 SNPs which were present only once in one of the isolates. Of those, 27,911 SNPs were found in isolates of the PA14 phylogroup and 21,811 SNPs in isolates of the PAO1 phylogroup.

As expected the sequence variation among the isolates was not random but isolates that belonged to the same phylogroup shared patterns of SNP profiles (Figure 17). The PA14 and PAO1 clonal complexes have diverged evolutionary [18, 59, 67] and I identified 463 positions where the entire 44 PA14-like isolates could be distinguished from the 55 PAO1-like isolates (strict phylogroup SNPs: present in all isolates of one phylogroup and in none of the other phylogroup). I found 8,410 additional SNPs, which were present in 100% of the either phylogroup with just one or two isolates of the other group harboring the respective SNP at the same position. In addition to these phylogroup-SNPs, 66,892 SNPs were found to be exclusively present in phylogenetically closely related isolates. These SNPs were more or less present in only one phylogenetic sub-group but were absent outside this group. For example, I found a particular SNP that was present in less than 44 of the PA14-like isolates and completely absent in entire PAO1-like isolates. Overall, 75,765 phylogroup/clonal SNPs were comprehensively classified by our approach. Of note, the majority of these clonal polymorphisms did not result in a change in the amino acid sequence of the encoding protein.

In addition to the 75,765 phylogroup/clonal SNPs and the 49,722 single SNPs (which occurred only in one isolate), a total of 34,122 SNPs were commonly identified across many isolates independent on the phylogenetic background (21.38% of the total SNPs in overall 3,567 genes).

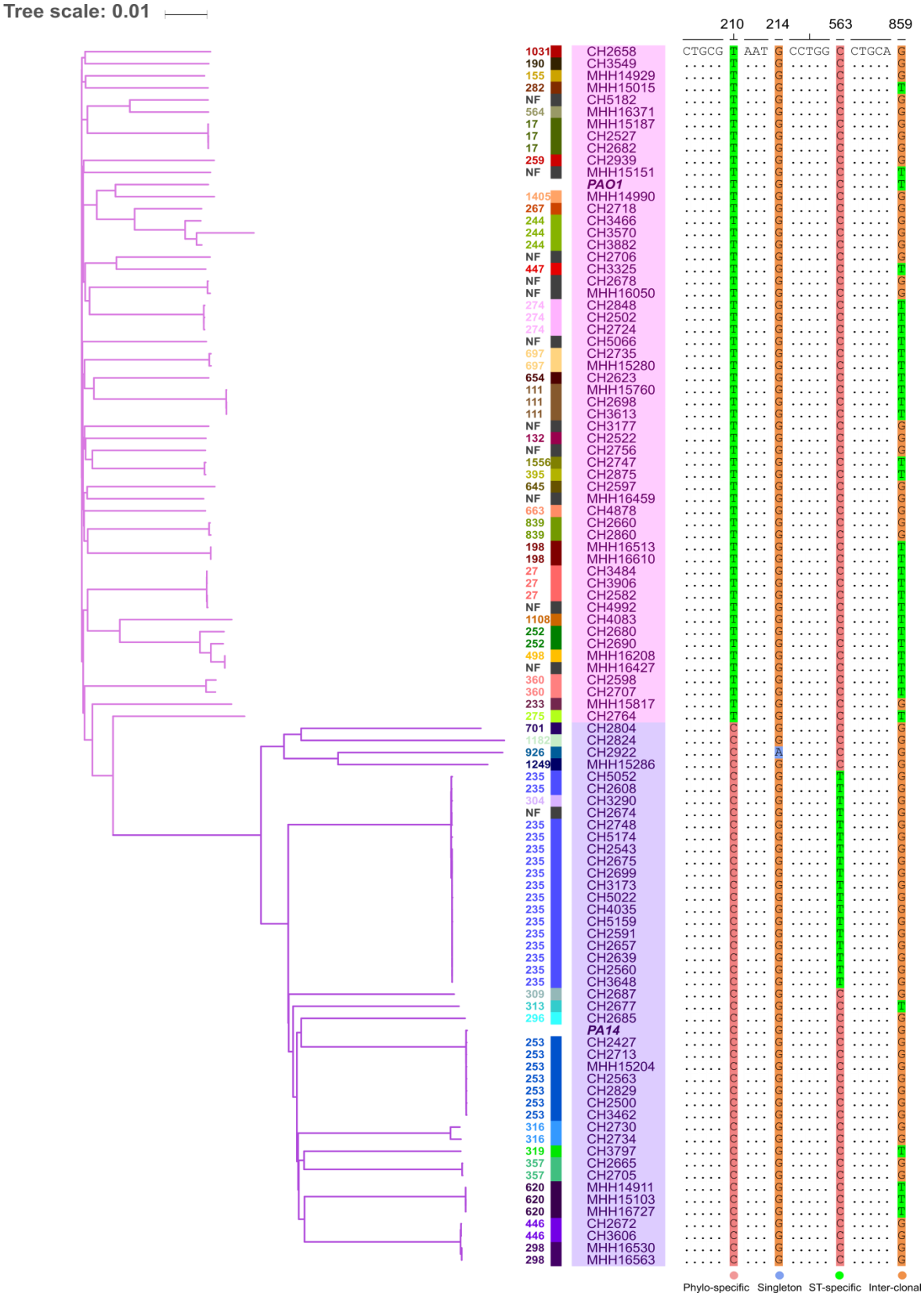

Tree scale: 0.01
210
214
563
859
CTGCG
AAT
CCTGG
CTGCA
CH2658
CH3549
MHH14929
MHH15015
CH5182
MHH16371
MHH15187
CH2527
CH2682
CH2939
MHH15151
PAO1
MHH14990
CH2718
CH3466
CH3570
CH3882
CH2706
CH3325
CH2678
MHH16050
CH2848
CH2502
CH2724
CH5066
CH2735
MHH15280
CH2623
MHH15760
CH2698
CH3613
CH3177
CH2522
CH2756
CH2747
CH2875
CH2597
MHH16459
CH4878
CH2660
CH2860
MHH16513
MHH16610
CH3484
CH3906
CH2582
CH4992
CH4083
CH2680
CH2690
MHH16208
MHH16427
CH2598
CH2707
MHH15817
CH2764
CH2804
CH2824
CH2922
MHH15286
CH5052
CH2608
CH3290
CH2674
CH2748
CH5174
CH2543
CH2675
CH2699
CH3173
CH5022
CH4035
CH5159
CH2591
CH2657
CH2639
CH2560
CH3648
CH2687
CH2677
CH2685
PA14
CH2427
CH2713
MHH15204
CH2563
CH2829
CH2500
CH3462
CH2730
CH2734
CH3797
CH2665
CH2705
MHH14911
MHH15103
MHH16727
CH2672
CH3606
MHH16530
MHH16563
Phylo-specific
Singleton
ST-specific
Inter-clonal

Figure 17: Classification of SNPs across isolates.

The phylogenetic tree reflects the ST-specific information in two major groups. Four positions of PA14_00190 are taken as an example to demonstrate categorization into different classes of SNPs. At position 210, the SNP is 100% clonal specific; at position 214, a single SNP is present only in isolate-CH2922; at position 563, a sequence type (ST) – specific SNP is present in ST235 isolates (PA14 phylogroup), and at position 859, the SNP is commonly identified in the clinical isolates and is found independent on the phylogenetic background.

Of these 34,122 SNPs, 27,590 SNPs (approximately 80%) were identified as synonymous (in overall 3,497 genes) whereas 6,532 SNPs (approximately 20%) were identified as non-synonymous (in overall 2,252 genes). In addition to the 75,765 phylogroup/clonal SNPs and the 49,722 single SNPs (which occurred only in one isolate), a total of 34,122 SNPs were commonly identified across many isolates independent on the phylogenetic background (21.38% of the total SNPs in overall 3,567 genes). Of these 34,122 SNPs, 27,590 SNPs (approximately 80%) were identified as synonymous (in overall 3,497 genes) whereas 6,532 SNPs (approximately 20%) were identified as non-synonymous (in overall 2,252 genes) (Figure 18). A list that ranks these 2,252 genes according to the relative frequency of non-synonymous mutations and normalization is provided in supplementary table. I also determined the frequency of mutations at identical nucleotide positions across the clinical isolates. The position-wise mutations were sorted based on the number of occurrences across the total number of isolates and presented in supplementary table.

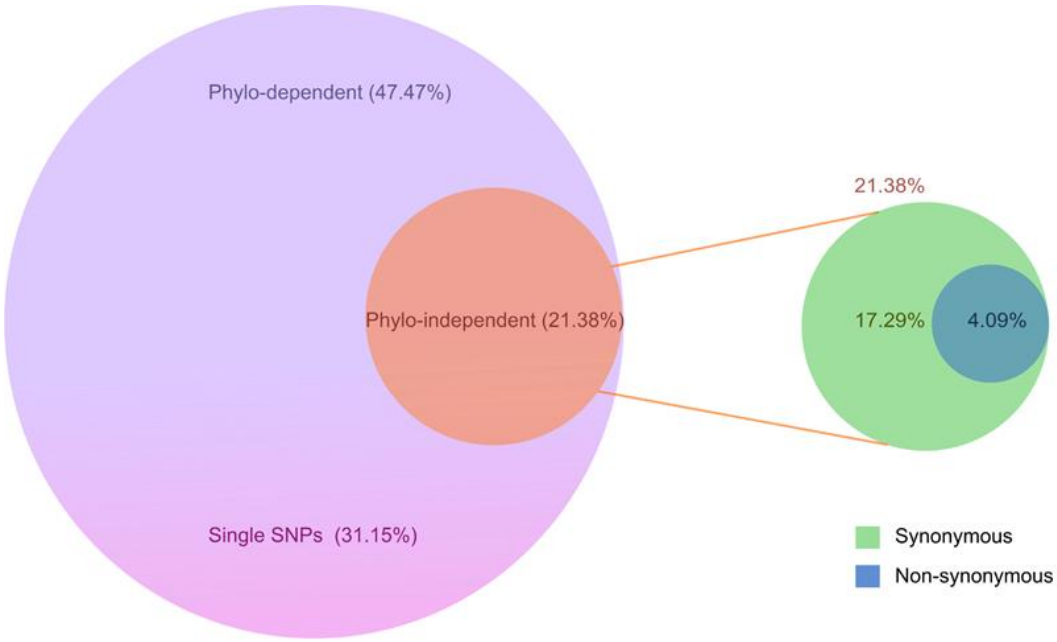


Figure 18: SNPs within the core genome of *P. aeruginosa*.

In a total of 159,609 SNPs, approximately 47.47% are phylogenetically related (either clonally or sequence type-specific) – highlighted in violet; 31.15% are present in only one of the isolates (single SNPs) – highlighted in pink and 21.38% are identified in the isolates independent on the phylogenetic background – highlighted in orange. The latter SNPs include 17% of synonymous mutations – highlighted in green and 4% of non-synonymous mutations – highlighted in blue.

3.2.5 dN/dS ratio as a measure of selective pressure

The ratio of nonsynonymous substitutions rates (dN) to synonymous substitutions rates (dS), dN/dS, remains one of the most common measures used to describe stabilizing selection. A ratio of dN/dS > 1 indicates positive selection, dN/dS = 1 neutral selection, and dN/dS < 1 purifying (or negative) selection [68, 69]. To determine the synonymous and nonsynonymous substitution rates and the selective evolutionary pressure, a pairwise comparison approach was applied on the core genes using SNAP program based on the NG86 method [70, 71]. I found that the mean pairwise dN/dS ratio for the 3,814 core genes was 0.14 (Figure 19). This suggests that the core

genome of *P. aeruginosa* is under purifying selection as a whole, which is in agreement with previous analyses [72, 73]. I found only six genes (*bfrG*, *fptB*, *napA*, PA14_11160, PA14_65950, and PA14_69250) that exhibited a dN/dS ratio between 1 and 2.

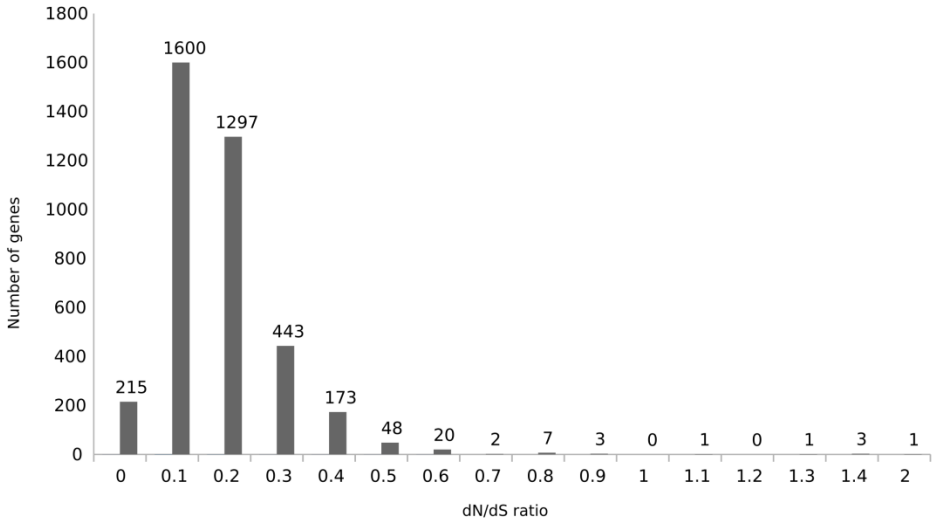


Figure 19: dN/dS ratio (omega values) for the overall 3,814 core genes.

The mean dN/dS ratio for the 3,814 core genes is 0.14. The ratio of almost all the core genes is between 0 to 1. There are only six genes that show the dN/dS ratio between 1 and 2.

3.3 Discussion

The advances of next generation sequencing opened up new frontiers in microbial genomics. Relatively inexpensive NGS technologies can produce large quantities of sequencing data, so that today genome analysis moves from comparing the sequence variation among two or few strains to the analysis of the genetic variation within species [52-57, 72, 73]. Deeper insights into the unique genetic makeup does not only allow to distinguish bacterial species from one another, but also determines their capacity to adapt to changing conditions and, potentially, to produce new species [74, 75]. However, what is lacking is a set of standardized analysis tools that can be used to describe a population-wide genomic diversity and to study how this relates to phenotype heterogeneity. There are several essential computational problems that need to be addressed before the information contained within genome sequences can be fully accessed [34, 76].

In order to describe the genetic variation in a given bacterial population, it is common practice to use a reference genome that acts as a general coordinate system and the mapping approach as the way of detecting genetic variations between a sample and the reference. In this study, I concentrated on directly aligning sequencing data without the need of a reference. Direct alignment approaches are commonly performed to uncover diversity among sequences; however, only at the single gene (or protein) level. I therefore identified in a first step the *P. aeruginosa* pan-genome, which gives information on the structure of the global microbial gene pool. The pan-genome of the 99 clinical isolates of this study compared well to previously established *P. aeruginosa* pan-genomes [59-62]. I supposed that the pan-genome information could be utilized to describe genetic variation by aligning orthologous sequences of the global gene pool in a gene-wise approach. I identified a large proportion of genes as *P. aeruginosa* core genes (3,814 genes) and generated a consensus sequence from the aligned orthologous sequences of all individual core genes. Any deviation from the consensus nucleotide at any position within the aligned genes gave information on SNPs at high accuracy and without the need to define cut-offs. The probability of the detection of true SNPs is highest when the orthologous sequences are identical in length. I thus applied our approach on the strict core genes of the fully sequenced 99 clinical *P. aeruginosa* isolates and ascertained SNPs across the population.

Importantly, in addition to the unambiguous detection of SNPs, their shared evolutionary origins can be assessed to infer phylogenetic relatedness. In the last decade multilocus sequence typing (MLST) has been used to study the molecular epidemiology of bacterial pathogens [77]. MLST measures genetic variation in a limited set of house-keeping genes that are usually PCR amplified and sequenced. The sequencing results are then used to assign individual isolates of a population to sequence types (ST) according to their unique allele profiles. Today, with next generation sequencing technologies information across the entire genome can be obtained and thus sequence types can be identified at a much higher degree of discrimination. Thereby the discrimination power increases with the number of core genes and the quality of the ascertained sequence variations [36, 78].

In this study, information of the genetic diversity across the entirety of the core genes of 99 clinical *P. aeruginosa* isolates provided the bases for unprecedentedly detailed information on phylogenetic relatedness between the individual isolates of the population. This information is also of importance when studying how genotypes are linked to phenotypes. Evolution does not only proceed through divergence of genes and ultimately phenotypes, but similar traits might also evolve convergently in unrelated isolates owing to similar selection pressures. In our study detailed knowledge on the full pattern of sequence variations among the 99 clinical *P. aeruginosa* isolates paved the way for the categorization of SNPs into those that are the basis for branching of the phylogenetic tree (phylogeny SNPs) and those that have been acquired independently, in separate lineages, and not through inheritance from a common ancestor. They might confer a selective advantage and are expected to be evolutionary short-lived. They should be found commonly in the clinical isolates but, in contrast to the phylogeny determining SNPs, they should not be fixed in the bacterial population. I identified nearly 47.47% of the total SNPs as phylogroup-dependent; whereas 21.38% were commonly found independent on the strain background. Approximately 31.15% of the total SNPs were singleton SNPs. The average number of singletons per isolate was 492. A large fraction of these singletons were identified in hypermutator strains (average 2000, n=15), however, interestingly, also both type strains PA14 and PAO1 exhibited far above average numbers of singletons – 811 and 1118 respectively. This indicates that purifying selection might not be at work, so that the lab strains have deviated quite substantially from the natural *P. aeruginosa* isolates.

We further classified the phylogenetically independent SNPs into synonymous (20%) and nonsynonymous mutations (80%) respectively and determined the ratio of the nonsynonymous substitution rate (dN) to the synonymous substitutions rate (dS), dN/dS for the 3,814 core genes. The resulting ratio of dN:dS = 0.14 indicates that the *P. aeruginosa* core genome is highly conserved across clinical isolates and in general not subject to adaptive evolution. In agreement, I found a sequence diversity at the single nucleotide polymorphism level of 0.04 (4×10^{-2}) in the strict core genes. This low sequence diversity corresponds to findings of previous studies, demonstrating that the core genes are highly conserved and differ by just a few SNPs from one isolate to the other [80].

Previous comparative genomic studies have demonstrated that the core genes of bacteria play important roles in niche adaptation and virulence especially in *P. aeruginosa* [66, 79]. However, the *P. aeruginosa* genome size varies from 5.2 to >7 Mbp demonstrating extensive accessory genome variability. Although many virulence determinants are generally a part of the core genome, novel accessory genomic sequences will continue to be detected [64, 81], and it has been suggested that the content of the accessory genome in *P. aeruginosa* determines environmental adaptability such as niche expansion [60, 67, 74]. A recent study further revealed that the intergenic mutations are more likely to be positively selected than coding mutations especially as this also enables essential genes to become targets of evolution in *P. aeruginosa* [82]. Furthermore, it seems that mainly mutations in master regulators impact on the bacterial phenotype. It was shown that the transcriptional profiles of some completely unrelated genotypes exhibited similar phenotypes; and highly similar genotypes exhibited substantially different transcriptional phenotypes, mainly due to inactivating mutations in global regulators [83]. Since inactivated global regulators often harbor indels or partial gene losses, many of them were not categorized in our pan-genome reconstruction as strict core genes.

In conclusion, I have shown that the core genome of the 99 clinical *P. aeruginosa* isolates analyzed in this study, is conserved and in general not under positive selection. However, in order to adequately address the question of how the *P. aeruginosa* genome shapes the bacterial phenotype, identification of sequence variations within the soft-core genes and the accessory genes and possibly also the intergenic regions will be inevitable. For this purpose the methods

established in this study, will have to be adjusted and re-fined so that full genomes can be evaluated at the population level. This will allow for detailed information on the full structure and dynamics of the *P. aeruginosa* SNP profile. The availability of this information will be an important step on our way to understand the causative link between genotypes and phenotypes. Inclusion of the *P. aeruginosa* genetic variation landscape in the web-based bactome database (https://bactome.helmholtz-hzi.de) [98] will provide the opportunity for data visualization in a graphical format that will make information on genetic diversity easily accessible.

3.4 MATERIALS AND METHODS

3.4.1 Bacterial strains and genomic sequencing

This study included 99 clinical isolates of *P. aeruginosa* that were sampled from various infection sites from patients at different hospitals across Germany. Genomic DNA was prepared from *P. aeruginosa* isolates using NEB Next Ultra Kit and sequenced on an Illumina MiSeq, generating 2x300-bp paired-end reads. Using a multiplexed protocol an average of 1,037,171 reads (range of 512,812 - 1,645,685) for each of the genomic libraries were obtained. On average, the isolates were sequenced with an estimated genomic coverage of 68-fold. The detailed information on the isolates and sequencing (reads) are provided in a supplementary table.

3.4.2 *De novo* assembly, annotation and generation of the pan-genome

Preprocessing such as the removal of adapter and bar code sequences were done using the FASTQ-MCF script included in the EA-UTILS (https://code.google.com/p/ea-utils/) [84]; and Karect was used for error correction [85]. The processed reads were assembled *de novo* using the A5-miseq pipeline and built-in scaffolder, SSPACE v3, was used to generate scaffolds from the assembled contigs, resulted in average of 40 scaffolds (13 to 192) [86]. The final assembled scaffolds were parsed to generate the gene annotation using Prokka v1.11 [87].

With the aim to assign genes to orthologous groups (gene families), I defined the genes present in all genomes (core genome) or in only a fraction of genomes (accessory genome) by blasting all-against-all using BLASTN. I selected hits with greater than 90% length and 90% sequence identity in the reciprocal set of homologs by a custom Perl script, as described previously [88, 89]. I then extracted gene sequences for all individual core genes from the corresponding genomes to define the orthologous group sequences. Based on the presence and absence of the gene information in each isolate, an orthologous matrix was created for statistical extrapolation analysis and saturation model. The formula used for the pan- and core- genome saturation model were $y = z + (a*x/(b+x))$ and $y = z - (a*x/(b+x))$ respectively.

3.4.3 Core Genome Multi Locus Sequence Typing (cgMLST)

The phylogenetic tree was created based on sequence variations within the 3,814 core genes. Custom Perl scripts were used to extract and concatenate the core gene sequences, resulting in one concatenated sequence per isolate. Phylogenetic distances between the strains were calculated using a *k*-mer approach, as described previously [90]. The sequences were split into 15-mers (and into 22-mers to construct the phylogenetic tree of 99 isolates plus all 52 reference genomes), which were then compared between the isolates. The resulting distance matrix was used to build a neighbor-joining tree in R using the ape package [91]. SRST tool and BIGSdb database were used to identify the sequence type (ST) information for each isolate [92, 93]. This information was supplemented and visualized in the phylogenetic tree using iTOL [94].

3.4.4 Consensus nucleotide sequence and SNP detection

The core gene orthologous groups (3,814 groups) were used to create a gene-wide consensus sequence. If genes within one orthologous group were identical in length, the orthologous sequences were directly aligned as a sequence matrix. A multiple sequence alignment was performed on the non-identical length group to identify the gaps by using clustal-omega [95]. These gaps were then filled with an "N" to have an identical number of nucleotide positions across all isolates before the orthologous sequences were aligned as a sequence matrix. A Perl script was developed to identify a position-wise proportionate distribution of the nucleotides across the genomes. The most frequently occurring nucleotide was considered consensus. If the most frequent nucleotide was ambiguous at a particular position, the selection of consensus was based on whether or not the nucleotide was present in a selected phylogenetic group (here PA14). In this case the consensus nucleotide was considered as a SNP (refer Figure 16).

3.4.5 SNP classification

Strain-wide information such as phylogroup and STs was additionally used to classify the phylogenetic-dependent and phylogenetically independent SNPs. The classification were checked by these conditions, (i) SNPs that occurred exclusively in only one phylogroup; (ii) SNPs that were fully present in one group plus the same SNP occurred in one or two isolates of

the other group; (iii) SNPs that were present in phylogenetically similar isolates (e.g. ST-specific) within one clonal complex and completely absent in the other group. Apart from this, single SNPs were detected if they were present in only one isolate. Furthermore, the substitutions in the amino acids were also detected and used to classify the phylogenetically independent SNPs into synonymous and non-synonymous mutations.

3.4.6 dN/dS ratio

dN/dS, the ratio of the rate of nonsynonymous substitutions (dN) to the rate of synonymous substitutions (dS) was calculated based on the modified Nei and Gojobori method by using the SNAP program. The sequences of the 3,814 genes from each isolate were used to calculate the dN/dS ratio.

3.4.7 Bactome database

Information on whether genes belong to the core or accessory genome and the consensus sequence of 3,014 core gene families is visualized in the Bactome database (https://bactome.helmholtz-hzi.de). The user can retrieve information by inserting a PA14 or PAO1 locus tag. Description of genetic diversity is currently based on genomic information of 99 isolates with the two references (in total 101 genomes). However, the database can be updated so that information also on the accessory genes can be obtained as well as information from a larger sample size of clinical isolates.

3.4.8 Nucleotide sequence accession number

All the short-read data are available at the National Center for Biotechnology Information Sequence Read Archive (http://www.ncbi.nlm.nih.gov/sra) under accession number SRP160898.

Outlook

Overall conclusion and outlook

Bacterial genome sequencing has become handy and multiple genome analyses of an individual species have revealed comprehensive genomic diversity [83]. Also RNA-sequencing has provided unbiased, accurate insight into the nature and dimension of a bacterial gene expression profile during the course of infection within the human host [47]. As these data have become available in large scale, it is mandate to have a multi-disciplinary approach to understand fully the genomic diversity as well as the global transcriptional profile of any bacterial pathogen.

With this thesis, two analytical pipelines were developed: the pan-genome based transcriptomic data analysis and the consensus nucleotide reference based SNP identification. The pan-genome based analysis is needed for identifying the gene functions that might be responsible for an acquired phenotype. The creation of pan-genomes in *E.coli* and *K. pneumoniae* facilitated the transcriptomic studies. While *E. coli* isolates exhibit a large general gene expression profile but they clearly express distinct accessory genomes. And the core transcriptome profiles of the clinical *K. pneumoniae* isolates were similar but distinct between groups sequence types (STs). With the consensus sequence approach, the clonal related SNPs and single SNPs were identified in the clinical *P. aeruginosa* isolates. I defined and identified the phylogenetically independent mutations that are occurring as recurrent combinations of patterns in many of the core genes but the core genome was under purifying selection as a whole. The pattern of inter-clonal nonsynonymous mutations composes a large repertoire of genes from both the core and accessory genome – whereas a more comprehensive search is needed for the patho-adaptive mutations. Thus, the identification of precise candidate patho-adaptive mutations/genes is crucial to understand the evolution at genetic level.

In conclusion, this thesis demonstrates the usefulness bioinformatic analytical pipelines in order to attain relevant information from the genomic and transcriptomic data. The identification of the impact of gene expression on the phenotype as well as the identification of the gene functions that might be responsible for an acquired phenotype was achieved by the pan-genome based analysis. Deeper insights were gained on the sequence variations of invariant *P. aeruginosa,* and by classifying SNPs, into clonal specific, single SNPs and phylogenetically independent SNPs; I present a novel and promising approach to identify the candidate patho-adaptive genes. The

proposed approach has the flexibility to incorporate the soft core genes. Full information on patho-adaptive mutations of the core and the soft core genome will promise to improve our understanding on the importance of pathogenicity-adaptive mutations in adaptive evolution.

References

1. Berger, B., Peng, J., & Singh, M. (2013). Computational solutions for omics data. Nature Reviews Genetics, 14(5), 333.
2. Loman, N. J., Pallen, M. J. (2015). Twenty years of bacterial genome sequencing. Nature Reviews Microbiology, 13(12), 787.
3. Watson, J. D., Crick, F. H. (1953). Molecular structure of nucleic acids. Nature, 171(4356), 737-738.
4. Sanger, F., Nicklen, S., Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. Proceedings of the national academy of sciences, 74(12), 5463-5467.
5. International Human Genome Sequencing Consortium. (2001). Initial sequencing and analysis of the human genome. Nature, 409(6822), 860.
6. Dayhoff, M. O. (1965). Atlas of protein sequence and structure. Silver Spring, Md: National Biomedical Research Foundation.
7. Fitch, W. M., Margoliash, E. (1967). Construction of phylogenetic trees. Science, 155(3760), 279-284.
8. Needleman, S. B., Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of molecular biology, 48(3), 443-453.
9. Smith, T. F., Waterman, M. S. (1981). Identification of common molecular subsequences. Journal of molecular biology, 147(1), 195-197.
10. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., Lipman, D. J. (1990). Basic local alignment search tool. Journal of molecular biology, 215(3), 403-410.
11. Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., et al. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic acids research, 25(17), 3389-3402.
12. Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., et al. (1995). Whole genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science, 269(5223), 496-512.
13. Blattner, F. R., Plunkett, G., Bloch, C. A., Perna, N. T., Burland, V., et al. (1997). The complete genome sequence of *Escherichia coli* K-12. Science, 277(5331), 1453-1462.

14. Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000). Gene Ontology: tool for the unification of biology. Nature genetics, 25(1), 25.
15. Kanehisa, M., Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. Nucleic acids research, 28(1), 27-30.
16. Brown, C. T., Hug, L. A., Thomas, B. C., Sharon, I., Castelle, C. J., et al. (2015). Unusual biology across a group comprising more than 15% of domain Bacteria. Nature, 523(7559), 208.
17. Stover, C. K., Pham, X. Q., Erwin, A. L., Mizoguchi, S. D., Warrener, P., et al. (2000). Complete genome sequence of *Pseudomonas aeruginosa* PAO1, an opportunistic pathogen. Nature, 406(6799), 959.
18. Lee, D. G., Urbach, J. M., Wu, G., Liberati, N. T., Feinbaum, R. L., et al. (2006). Genomic analysis reveals that *Pseudomonas aeruginosa* virulence is combinatorial. Genome biology, 7(10), R90.
19. Roy, P. H., Tetu, S. G., Larouche, A., Elbourne, L., Tremblay, S., et al. (2010). Complete genome sequence of the multiresistant taxonomic outlier *Pseudomonas aeruginosa* PA7. PloS one, 5(1), e8842.
20. Winsor, G. L., Lo, R., Sui, S. J. H., Ung, K. S., Huang, S., et al. (2005). *Pseudomonas aeruginosa* Genome Database and PseudoCAP: facilitating community-based, continually updated, genome annotation. Nucleic acids research, 33(suppl_1), D338-D343.
21. Cochrane, G., Karsch-Mizrachi, I., Takagi, T., Sequence Database Collaboration, I. N. (2015). The international nucleotide sequence database collaboration. Nucleic acids research, 44(D1), D48-D50.
22. Toribio, A. L., Alako, B., Amid, C., Cerdeño-Tarrága, A., Clarke, L., et al. (2016). European nucleotide archive in 2016. Nucleic acids research, 45(D1), D32-D36.
23. Mashima, J., Kodama, Y., Fujisawa, T., Katayama, T., Okuda, Y., et al. (2016). DNA Data Bank of Japan. Nucleic acids research, gkw1001.
24. Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., et al. (2017). GenBank. Nucleic acids research. 45(Database issue), D37–D42. http://doi.org/10.1093/nar/gkw1070.

25. Stephens, Z. D., Lee, S. Y., Faghri, F., Campbell, R. H., Zhai, C., et al. (2015). Big data: astronomical or genomical?. PLoS biology, 13(7), e1002195.
26. Chojnacki, S., Cowley, A., Lee, J., Foix, A., Lopez, R. (2017). Programmatic access to bioinformatics tools from EMBL-EBI update: 2017. Nucleic acids research, 45(Web Server issue), W550.
27. Galperin, M. Y., Fernández-Suárez, X. M., Rigden, D. J. (2017). The 24th annual Nucleic Acids Research database issue: a look back and upcoming changes. Nucleic acids research, 45(D1), D1-D11.
28. Schadt, E. E. (2012). The changing privacy landscape in the era of big data. Molecular systems biology, 8(1), 612.
29. Tettelin, H., Masignani, V., Cieslewicz, M. J., Donati, C., Medini, D., et al. (2005). Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome". Proceedings of the National Academy of Sciences of the United States of America, 102(39), 13950-13955.
30. Vernikos, G., Medini, D., Riley, D. R., Tettelin, H. (2015). Ten years of pan-genome analyses. Current opinion in microbiology, 23, 148-154.
31. Rouli, L., Merhej, V., Fournier, P. E., Raoult, D. (2015). The bacterial pangenome as a new tool for analysing pathogenic bacteria. New microbes and new infections, 7, 72-85.
32. Li, L., Stoeckert, C. J., Roos, D. S. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome research, 13(9), 2178-2189.
33. Moreno-Hagelsieb, G., Latimer, K. (2007). Choosing BLAST options for better detection of orthologs as reciprocal best hits. Bioinformatics, 24(3), 319-324.
34. Computational Pan-Genomics Consortium. (2016). Computational pan-genomics: status, promises and challenges. Briefings in bioinformatics.
35. Medini, D., Serruto, D., Parkhill, J., Relman, D. A., Donati, C., et al. (2008). Microbiology in the post-genomic era. Nature Reviews Microbiology, 6(6), 419.
36. Maiden, M. C., Van Rensburg, M. J. J., Bray, J. E., Earle, S. G., Ford, S. A., et al. (2013). MLST revisited: the gene-by-gene approach to bacterial genomics. Nature Reviews Microbiology, 11(10), 728.

37. Hurgobin, B., Edwards, D. (2017). SNP discovery using a pangenome: has the single reference approach become obsolete?. Biology, 6(1), 21.
38. Donnenberg, M. S., and Whittam, T. S. (2001). Pathogenesis and evolution of virulence in enteropathogenic and enterohemorrhagic *Escherichia coli*. The Journal of clinical investigation, 107(5), 539-548.
39. Fukiya, S., Mizoguchi, H., Tobe, T., & Mori, H. (2004). Extensive genomic diversity in pathogenic *Escherichia coli* and *Shigella* strains revealed by comparative genomic hybridization microarray. Journal of bacteriology, 186(12), 3911-3921.
40. Zhang, Y., and Lin, K. (2012). A phylogenomic analysis of *Escherichia coli/Shigella* group: implications of genomic features associated with pathogenicity and ecological adaptation. BMC evolutionary biology, 12(1), 174.
41. Köhler, C. D., & Dobrindt, U. (2011). What defines extraintestinal pathogenic *Escherichia coli*?. International Journal of Medical Microbiology, 301(8), 642-647.
42. Narciso, A., Nunes, F., Amores, T., Lito, L., et al. (2012). Persistence of uropathogenic *Escherichia coli* strains in the host for long periods of time: relationship between phylogenetic groups and virulence factors. European journal of clinical microbiology and infectious diseases, 31(6), 1211-1217.
43. Sabaté, M., Moreno, E., Pérez, T., Andreu, A., et al. (2006). Pathogenicity island markers in commensal and uropathogenic *Escherichia coli* isolates. Clinical Microbiology and Infection, 12(9), 880-886.
44. Ramos, N. L., Sekikubo, M., Dzung, D. T. N., Kosnopfel, C., et al. (2012). Uropathogenic *Escherichia coli* isolates from pregnant women in different countries. Journal of clinical microbiology, 50(11), 3569-3574.
45. Holt, K. E., Wertheim, H., Zadoks, R. N., Baker, S., Whitehouse, C. A., et al. (2015). Genomic analysis of diversity, population structure, virulence, and antimicrobial resistance in *Klebsiella pneumoniae*, an urgent threat to public health. Proceedings of the National Academy of Sciences, 112(27), E3574-E3581.
46. Sridhar, J., Narmada, S. R., Sabarinathan, R., Ou, H. Y., et al. (2010). sRNAscanner: a computational tool for intergenic small RNA detection in bacterial genomes. Plos one, 5(8), e11970.

47. Wang, Z., Gerstein, M., Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. Nature reviews genetics, 10(1), 57.
48. Sokurenko, E. V., Hasty, D. L., Dykhuizen, D. E. (1999). Pathoadaptive mutations: gene loss and variation in bacterial pathogens. Trends in microbiology, 7(5), 191-195.
49. Sokurenko, E. V., Gomulkiewicz, R., Dykhuizen, D. E. (2006). Source–sink dynamics of virulence evolution. Nature Reviews Microbiology, 4(7), 548.
50. Li, H., Ruan, J., Durbin, R. (2008). Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome research, gr-078212.
51. Langmead, B., Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. Nature methods, 9(4), 357.
52. Smith, E. E., Buckley, D. G., Wu, Z., Saenphimmachak, C., Hoffman, L. R., et al. (2006). Genetic adaptation by *Pseudomonas aeruginosa* to the airways of cystic fibrosis patients. Proceedings of the National Academy of Sciences, 103(22), 8487-8492.
53. Oliver, A., Cantón, R., Campo, P., Baquero, F., Blázquez, J. (2000). High frequency of hypermutable *Pseudomonas aeruginosa* in cystic fibrosis lung infection. Science, 288(5469), 1251-1253.
54. Bragonzi, A., Paroni, M., Nonis, A., Cramer, N., Montanari, S., et al. (2009). *Pseudomonas aeruginosa* microevolution during cystic fibrosis lung infection establishes clones with adapted virulence. American journal of respiratory and critical care medicine, 180(2), 138-145.
55. Cramer, N., Klockgether, J., Wrasman, K., Schmidt, M., Davenport, C. F., et al. (2011). Microevolution of the major common *Pseudomonas aeruginosa* clones C and PA14 in cystic fibrosis lungs. Environmental microbiology, 13(7), 1690-1704.
56. Folkesson, A., Jelsbak, L., Yang, L., Johansen, H. K., Ciofu, O., et al. (2012). Adaptation of *Pseudomonas aeruginosa* to the cystic fibrosis airway: an evolutionary perspective. Nature Reviews Microbiology, 10(12), 841.
57. Marvig, R. L., Sommer, L. M., Molin, S., Johansen, H. K. (2015). Convergent evolution and adaptation of *Pseudomonas aeruginosa* within patients with cystic fibrosis. Nature genetics, 47(1), 57.

58. Medini, D., Donati, C., Tettelin, H., Masignani, V., Rappuoli, R. (2005). The microbial pan-genome. Current opinion in genetics & development, 15(6), 589-594.
59. Hilker, R., Munder, A., Klockgether, J., Losada, P. M., Chouvarine, et al. (2015). Interclonal gradient of virulence in the *Pseudomonas aeruginosa* pangenome from disease and environment. Environmental microbiology, 17(1), 29-46.
60. Mathee, K., Narasimhan, G., Valdes, C., Qiu, X., Matewish, J. M., et al. (2008). Dynamics of *Pseudomonas aeruginosa* genome evolution. Proceedings of the National Academy of Sciences, 105(8), 3100-3105.
61. Klockgether, J., Cramer, N., Wiehlmann, L., Davenport, C. F., Tümmler, B. (2011). *Pseudomonas aeruginosa* genomic structure and diversity. Frontiers in microbiology, 2, 150.
62. Valot, B., Guyeux, C., Rolland, J. Y., Mazouzi, K., Bertrand, X., et al. (2015). What it takes to be a *Pseudomonas aeruginosa*? The core genome of the opportunistic pathogen updated. PLoS One, 10(5), e0126468.
63. Curran, B., Jonas, D., Grundmann, H., Pitt, T., Dowson, C. G. (2004). Development of a multilocus sequence typing scheme for the opportunistic pathogen *Pseudomonas aeruginosa*. Journal of clinical microbiology, 42(12), 5644-5649.
64. Wiehlmann, L., Wagner, G., Cramer, N., Siebert, B., Gudowius, P., et al. (2007). Population structure of *Pseudomonas aeruginosa*. Proceedings of the National Academy of Sciences, 104(19), 8101-8106.
65. Freschi, L., Jeukens, J., Kukavica-Ibrulj, I., Boyle, B., Dupont, M. J., et al. (2015). Clinical utilization of genomics data produced by the international *Pseudomonas aeruginosa* consortium. Frontiers in microbiology, 6, 1036.
66. Spencer, D. H., Kas, A., Smith, E. E., Raymond, C. K., Sims, E. H., et al. (2003). Whole-genome sequence variation among multiple isolates of *Pseudomonas aeruginosa*. Journal of bacteriology, 185(4), 1316-1325.
67. Fischer, S., Klockgether, J., Morán Losada, P., Chouvarine, P., Cramer, N., et al. (2016). Intraclonal genome diversity of the major *Pseudomonas aeruginosa* clones C and PA14. Environmental microbiology reports, 8(2), 227-234.
68. Yang, Z., Bielawski, J. P. (2000). Statistical methods for detecting molecular adaptation. Trends in ecology & evolution, 15(12), 496-503.

69. Rocha, E. P., Smith, J. M., Hurst, L. D., Holden, M. T., Cooper, J. E., et al. (2006). Comparisons of d**N**/d**S** are time dependent for closely related bacterial genomes. Journal of theoretical biology, 239(2), 226-235.
70. Korber, B. (2000). HIV signature and sequence variation analysis. Computational analysis of HIV molecular sequences, 4, 55-72.
71. Nei, M., Gojobori, T. (1986). Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Molecular biology and evolution, 3(5), 418-426.
72. Yang, L., Jelsbak, L., Marvig, R. L., Damkiær, S., Workman, C. T., et al. (2011). Evolutionary dynamics of bacteria in a human host environment. Proceedings of the National Academy of Sciences, 108(18), 7481-7486.
73. Mosquera-Rendón, J., Rada-Bravo, A. M., Cárdenas-Brito, S., Corredor, M., Restrepo-Pineda, E., et al. (2016). Pangenome-wide and molecular evolution analyses of the *Pseudomonas aeruginosa* species. BMC genomics, 17(1), 45.
74. Silby, M. W., Winstanley, C., Godfrey, S. A., Levy, S. B., Jackson, R. W. (2011). *Pseudomonas* genomes: diverse and adaptable. FEMS microbiology reviews, 35(4), 652-680.
75. Caputo, A., Merhej, V., Georgiades, K., Fournier, P. E., Croce, O., Robert, C., et al. (2015). Pan-genomic analysis to redefine species and subspecies based on quantum discontinuous variation: the Klebsiella paradigm. Biology direct, 10(1), 55.
76. Xiao, J., Zhang, Z., Wu, J., Yu, J. (2015). A brief review of software tools for pangenomics. Genomics, proteomics & bioinformatics, 13(1), 73-76.
77. Maiden, M. C., Bygraves, J. A., Feil, E., Morelli, G., Russell, J. E., et al. (1998). Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. Proceedings of the National Academy of Sciences, 95(6), 3140-3145.
78. Leopold, S. R., Goering, R. V., Witten, A., Harmsen, D., Mellmann, A. (2014). Bacterial whole-genome sequencing revisited: portable, scalable, and standardized analysis for typing and detection of virulence and antibiotic resistance genes. Journal of clinical microbiology, 52(7), 2365-2370.

79. Wolfgang, M. C., Kulasekara, B. R., Liang, X., Boyd, D., Wu, K., et al. (2003). Conservation of genome content and virulence determinants among clinical and environmental isolates of *Pseudomonas aeruginosa*. Proceedings of the National Academy of Sciences, 100(14), 8484-8489.
80. Dötsch, A., Klawonn, F., Jarek, M., Scharfe, M., Blöcker, H., et al. (2010). Evolutionary conservation of essential and highly expressed genes in *Pseudomonas aeruginosa*. BMC genomics, 11(1), 234.
81. Pohl, S., Klockgether, J., Eckweiler, D., Khaledi, A., Schniederjans, M., et al. (2014). The extensive set of accessory *Pseudomonas aeruginosa* genomic components. FEMS microbiology letters, 356(2), 235-241.
82. Khademi, H., Jelsbak, L. (2017). Host adaptation mediated by intergenic evolution in a bacterial pathogen. bioRxiv, 236000.
83. Dötsch, A., Schniederjans, M., Khaledi, A., Hornischer, K., Schulz, S., et al. (2015). The *Pseudomonas aeruginosa* transcriptional landscape is shaped by environmental heterogeneity and genetic variation. MBio, 6(4), e00749-15.
84. Aronesty, E. (2013). Comparison of sequencing utility programs. The Open Bioinformatics Journal, 7(1).
85. Allam, A., Kalnis, P., Solovyev, V. (2015). Karect: accurate correction of substitution, insertion and deletion errors for next-generation sequencing data. Bioinformatics, 31(21), 3421-3428.
86. Coil, D., Jospin, G., Darling, A. E. (2014). A5-miseq: an updated pipeline to assemble microbial genomes from Illumina MiSeq data. Bioinformatics, 31(4), 587-589.
87. Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. Bioinformatics, 30(14), 2068-2069.
88. Bielecki, P., Muthukumarasamy, U., Eckweiler, D., Bielecka, A., et al. (2014). In vivo mRNA profiling of uropathogenic *Escherichia coli* from diverse phylogroups reveals common and group-specific gene expression profiles. MBio, 5(4), e01075-14.
89. Bruchmann, S., Muthukumarasamy, U., Pohl, S., Preusse, M., et al. (2015). Deep transcriptome profiling of clinical *Klebsiella pneumoniae* isolates reveals strain and sequence type-specific adaptation. Environmental microbiology, 17(11), 4690-4710.

90. Leekitcharoenphon, P., Nielsen, E. M., Kaas, R. S., Lund, O., Aarestrup, F. M. (2014). Evaluation of whole genome sequencing for outbreak detection of *Salmonella enterica*. PloS one, 9(2), e87991.
91. Paradis, E., Claude, J., Strimmer, K. (2004). APE: analyses of phylogenetics and evolution in R language. Bioinformatics, 20(2), 289-290.
92. Inouye, M., Dashnow, H., Raven, L. A., Schultz, M. B., Pope, B. J., et al. (2014). SRST2: Rapid genomic surveillance for public health and hospital microbiology labs. Genome medicine, 6(11), 90.
93. Jolley, K. A., Maiden, M. C. (2010). BIGSdb: scalable analysis of bacterial genome variation at the population level. BMC bioinformatics, 11(1), 595.
94. Letunic, I., Bork, P. (2006). Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. Bioinformatics, 23(1), 127-128.
95. Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., et al. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Molecular systems biology, 7(1), 539.
96. Anders, S., Huber, W. (2010). Differential expression analysis for sequence count data. Genome biology, 11(10), R106.
97. Dötsch, A., Eckweiler, D., Schniederjans, M., Zimmermann, A., Jensen, V., et al. (2012). The *Pseudomonas aeruginosa* transcriptome in planktonic cultures and static biofilms using RNA sequencing. PloS one, 7(2), e31092.
98. Hornischer, K., Khaledi, A., Pohl, S., Schniederjans, M., Pezoldt, L., et al. (2018). BACTOME - A reference database to explore the sequence- and gene expression- variation landscape of *Pseudomonas aeruginosa* clinical isolates. Nucleic acids research, gky895, https://doi.org/10.1093/nar/gky895

Appendix

List of 12,331 non-redundant genes used in *E. coli* pan-genome

(a) The data set comprises the gene ID, the number of orthologous, the gene name and product name if applicable and the normalized read counts per kilo base values for each gene in each of the 21 UTI isolates. (b) list of non-redundant gene IDs and their respective ortholog IDs (with sequence identity and length).

http://mbio.asm.org/content/5/4/e01075-14/DC9/embed/inline-supplementary-material-9.xls

List of 2,589 genes commonly expressed in all 21 UTI *E. coli* isolates

http://mbio.asm.org/content/5/4/e01075-14/DC2/embed/inline-supplementary-material-2.pdf

List of genes that are expressed specifically in the phylogroup A/B1 and B2

http://mbio.asm.org/content/5/4/e01075-14/DC7/embed/inline-supplementary-material-7.pdf

Bactome database

The visualization of the genetic diversity at the single nucleotide level of 3.014 *P. aeruginosa* core genes together with the consensus sequence is provided in the Bactome database.

Accessible through,

https://bactome.helmholtz-hzi.de.

Please check the tutorial link here

https://bactome.helmholtz-hzi.de/tutorial.html

Supplementary information

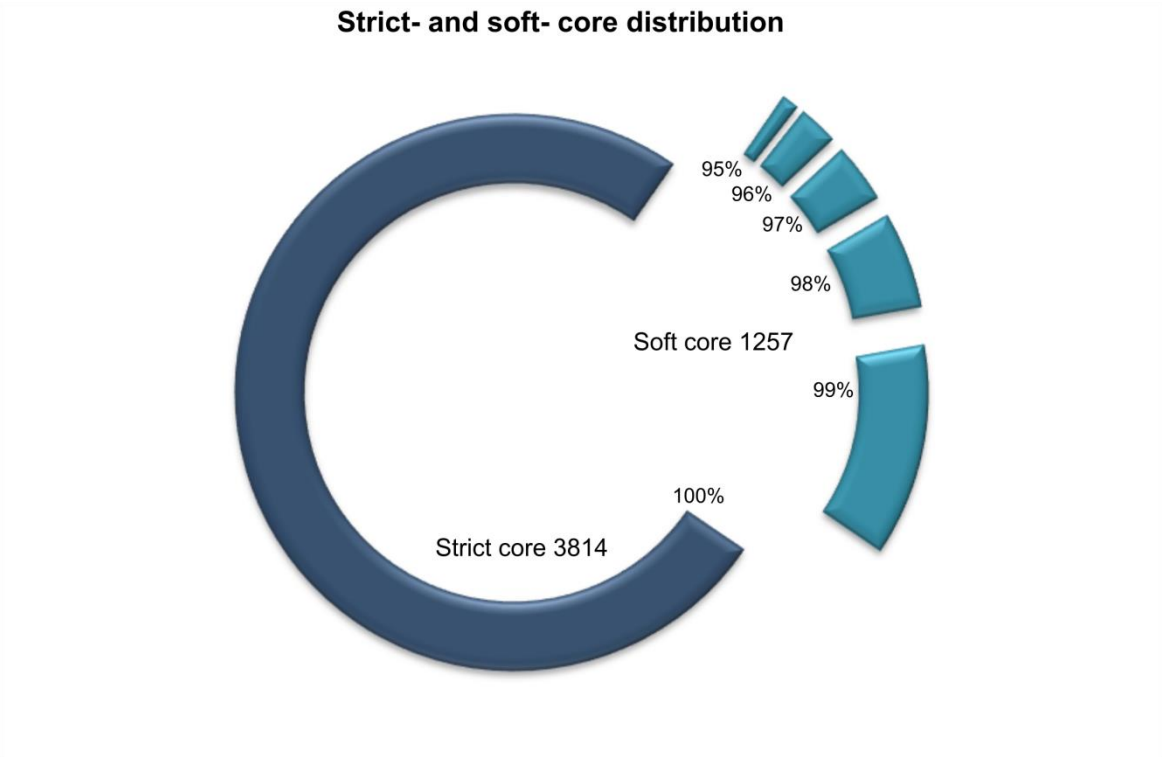


Figure S1: Distribution of the core and soft core genes in the group of 101 *P. aeruginosa* isolates.

In addition to the 3,814 core gene orthologous groups present in 100% of genomes, 1,257 soft core genes are identified in 99% (626), 98% (286), 97% (175), 96% (118) and 95% (52) of the genomes respectively.

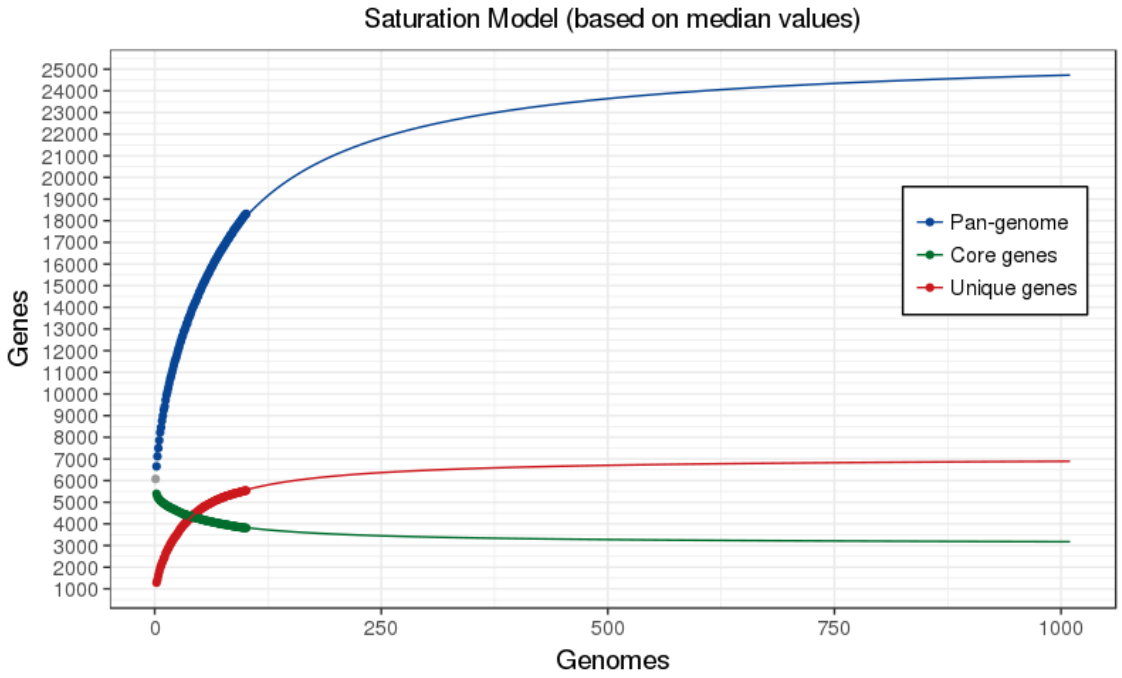


Figure S2: Saturation model of the pan-genome based on median values.

The saturation point for the pan- (blue), core- (green) and singleton (red) genomes based on the exponential expansion of presence and absence of gene is predicted by saturation model. It estimates that with the information of 685 genomes 95% saturation could be reached which would correspond to a pan-genome size of 25,882 genes including 3,062 core genes and 15,765 accessory genes, of which 7,055 would be singletons.

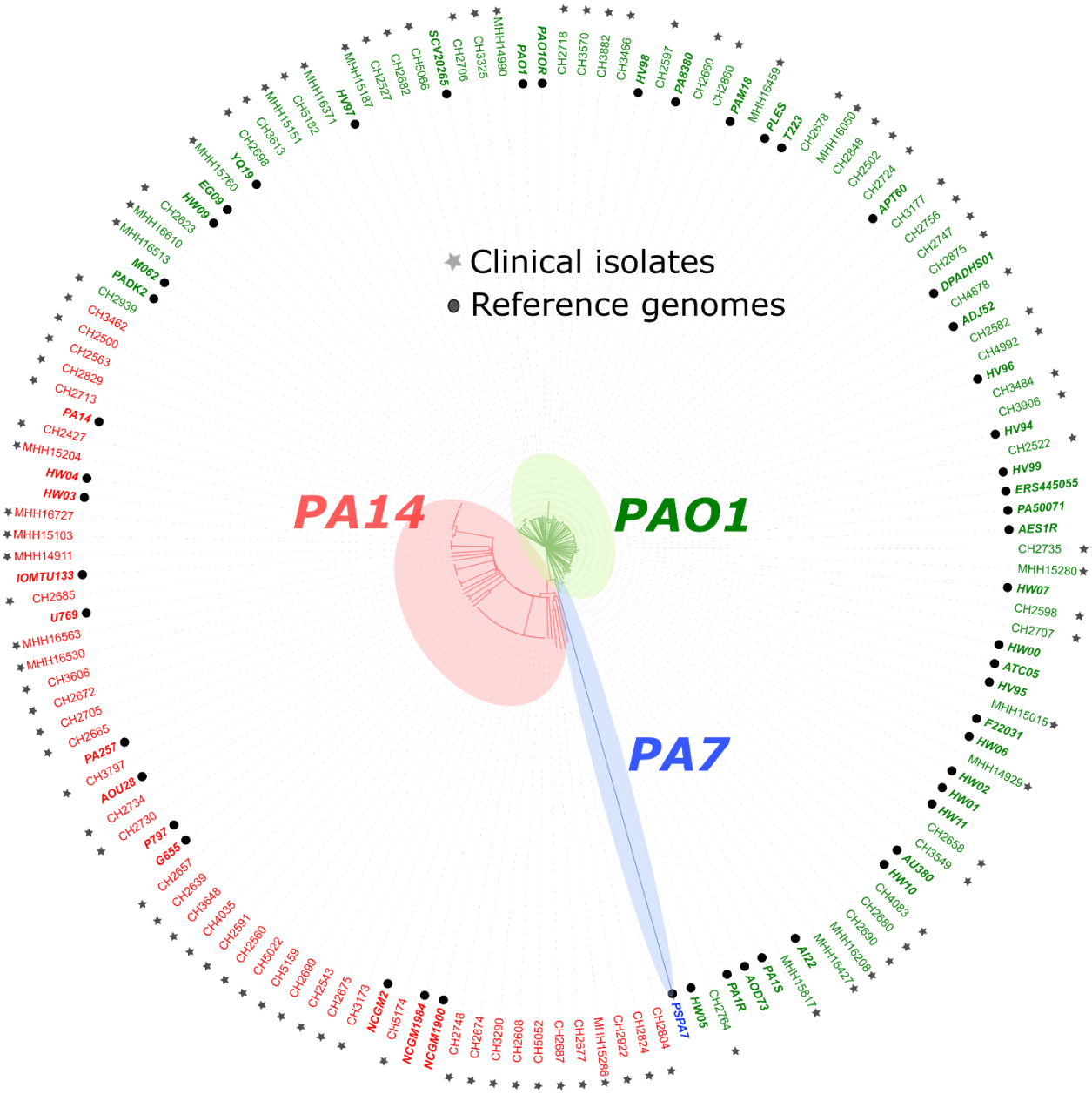


Figure S3: The phylogenetic distribution of 99 clinical isolates and 52 fully sequenced public reference genomes of *P. aeruginosa* as on 2016 (supplementary table S3). The current collection of clinical isolates is broadly distributed and comparable to the phylogenetic diversity of 52 previously sequenced *P. aeruginosa* genomes.

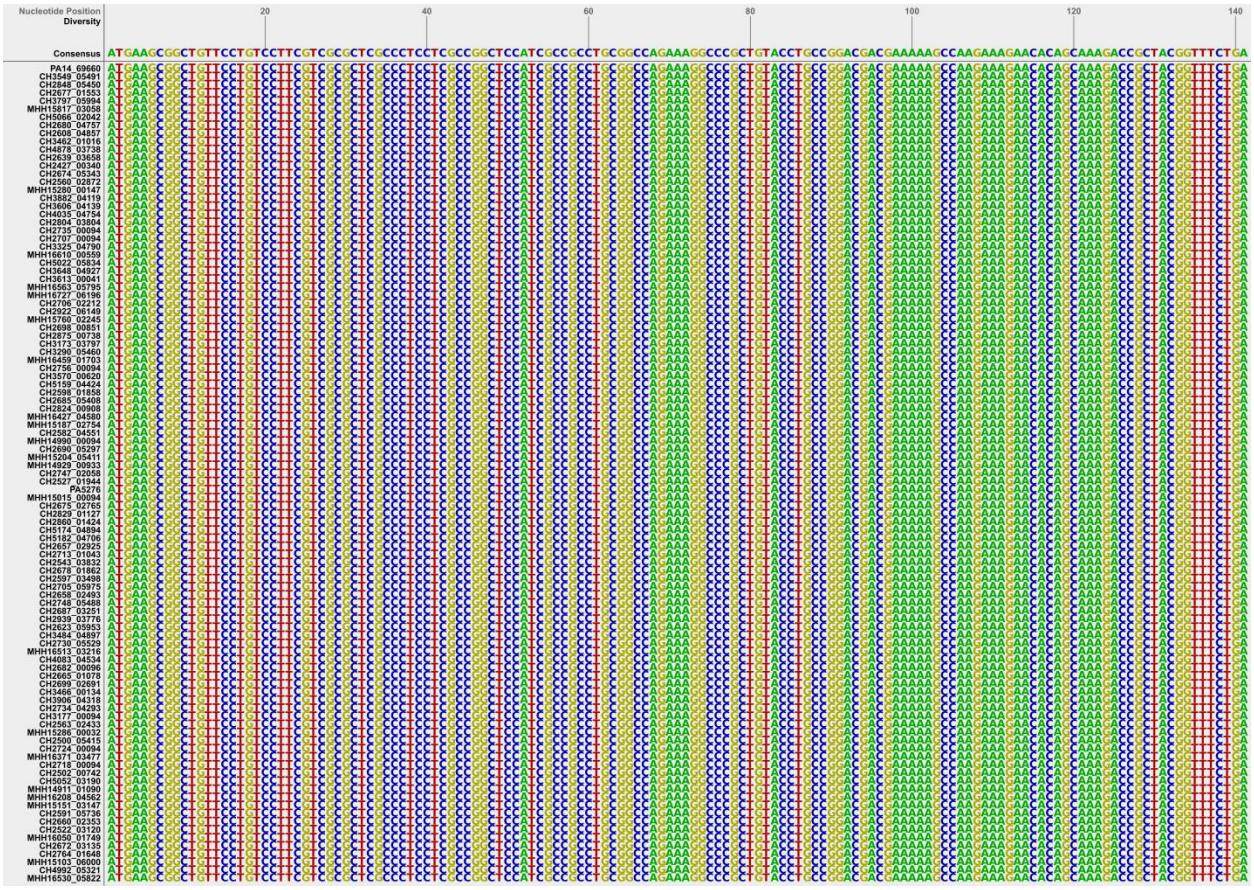


Figure S4: Consensus sequence and nucleotide diversity among clinical isolates visualized and stored in the Bactome database (https://bactome.helmholtz-hzi.de).

Table S1: Clinical *P. aeruginosa* isolates and infection information.

ID	Origin	Infection site/ sample	Type of infection
CH2427	Charité Berlin	other	other
CH2500	Charité Berlin	tracheobronchial secretion	respiratory tract
CH2502	Charité Berlin	nd/other	nd/other
CH2522	Charité Berlin	rectal smear	rectal infection
CH2527	Charité Berlin	groin/perineal	nd/other
CH2543	Charité Berlin	nasopharyngeal	respiratory tract
CH2560	Charité Berlin	wound swab	wound infection
CH2563	Charité Berlin	other	other
CH2582	Charité Berlin	swab	nd/other
CH2591	Charité Berlin	urinary tract	urinary tract
CH2597	Charité Berlin	nd/other	nd/other
CH2598	Charité Berlin	nd/other	nd/other
CH2608	Charité Berlin	wound swab	wound infection
CH2623	Charité Berlin	nd/other	nd/other
CH2639	Charité Berlin	tracheobronchial secretion	respiratory tract
CH2657	Charité Berlin	urinary tract	urinary tract
CH2658	Charité Berlin	nd/other	nd/other
CH2660	Charité Berlin	nd/other	nd/other
CH2665	Charité Berlin	tracheobronchial secretion	respiratory tract
CH2672	Charité Berlin	nd/other	nd/other
CH2674	Charité Berlin	tracheobronchial secretion	respiratory tract
CH2675	Charité Berlin	tracheobronchial secretion	respiratory tract
CH2677	Charité Berlin	urinary tract	urinary tract
CH2678	Charité Berlin	sputum CF	respiratory tract
CH2680	Charité Berlin	nd/other	nd/other
CH2682	Charité Berlin	nd/other	nd/other
CH2685	Charité Berlin	nd/other	nd/other
CH2687	Charité Berlin	urinary tract	urinary tract
CH2690	Charité Berlin	nd/other	nd/other
CH2698	Charité Berlin	nd/other	nd/other
CH2699	Charité Berlin	nd/other	nd/other
CH2705	Charité Berlin	rectal smear	rectal infection
CH2706	Charité Berlin	rectal smear	rectal infection
CH2707	Charité Berlin	urinary tract	urinary tract
CH2713	Charité Berlin	tracheobronchial secretion	respiratory tract
CH2718	Charité Berlin	nd/other	nd/other
CH2724	Charité Berlin	nd/other	nd/other
CH2730	Charité Berlin	nd/other	nd/other
CH2734	Charité Berlin	tracheobronchial secretion	respiratory tract
CH2735	Charité Berlin	tracheobronchial secretion	respiratory tract
CH2747	Charité Berlin	tracheobronchial secretion	respiratory tract
CH2748	Charité Berlin	tracheobronchial secretion	respiratory tract
CH2756	Charité Berlin	nd/other	nd/other

CH2764	Charité Berlin	wound swab	wound infection
CH2804	Charité Berlin	sputum CF	respiratory tract
CH2824	Charité Berlin	tracheobronchial secretion	respiratory tract
CH2829	Charité Berlin	rectal smear	rectal infection
CH2848	Charité Berlin	nd/other	nd/other
CH2860	Charité Berlin	nd/other	nd/other
CH2875	Charité Berlin	nd/other	nd/other
CH2922	Charité Berlin	tracheobronchial secretion	respiratory tract
CH2939	Charité Berlin	urinary tract	urinary tract
CH3173	Charité Berlin	rectal smear	rectal infection
CH3177	Charité Berlin	nd/other	nd/other
CH3290	Charité Berlin	sputum	respiratory tract
CH3325	Charité Berlin	nd/other	nd/other
CH3462	Charité Berlin	rectal smear	rectal infection
CH3466	Charité Berlin	blood	sepsis
CH3484	Charité Berlin	blood	sepsis
CH3549	Charité Berlin	sputum	respiratory tract
CH3570	Charité Berlin	nd/other	nd/other
CH3606	Charité Berlin	urinary tract	urinary tract
CH3613	Charité Berlin	nd/other	nd/other
CH3648	Charité Berlin	tracheobronchial secretion	respiratory tract
CH3797	Charité Berlin	stool	rectal infection
CH3882	Charité Berlin	urinary tract	urinary tract
CH3906	Charité Berlin	throat swab	respiratory tract
CH4035	Charité Berlin	tracheobronchial secretion	respiratory tract
CH4083	Charité Berlin	tracheobronchial secretion	respiratory tract
CH4878	Charité Berlin	tracheobronchial secretion	respiratory tract
CH4992	Charité Berlin	tracheobronchial secretion	respiratory tract
CH5022	Charité Berlin	nd/other	nd/other
CH5052	Charité Berlin	urinary tract	urinary tract
CH5066	Charité Berlin	nd/other	nd/other
CH5159	Charité Berlin	nd/other	nd/other
CH5174	Charité Berlin	nd/other	nd/other
CH5182	Charité Berlin	nd/other	nd/other
MHH14911	Hannover Medical School	wound swab	wound infection
MHH14929	Hannover Medical School	drain fluid	wound infection
MHH14990	Hannover Medical School	wound swab	wound infection
MHH15015	Hannover Medical School	tracheobronchial secretion	respiratory tract
MHH15103	Hannover Medical School	wound swab	wound infection
MHH15151	Hannover Medical School	tracheobronchial secretion	respiratory tract
MHH15187	Hannover Medical School	tracheobronchial secretion	respiratory tract
MHH15204	Hannover Medical School	nasopharyngeal	respiratory tract
MHH15280	Hannover Medical School	rectal smear	rectal infection
MHH15286	Hannover Medical School	tracheobronchial secretion	respiratory tract
MHH15760	Hannover Medical School	other	other
MHH15817	Hannover Medical School	blood	sepsis
MHH16050	Hannover Medical School	nd/other	nd/other
MHH16208	Hannover Medical School	wound swab	wound infection

MHH16371	Hannover Medical School	rectal smear	rectal infection
MHH16427	Hannover Medical School	wound swab	wound infection
MHH16459	Hannover Medical School	nasopharyngeal	respiratory tract
MHH16513	Hannover Medical School	nasopharyngeal	respiratory tract
MHH16530	Hannover Medical School	CF lung recipient	respiratory tract
MHH16563	Hannover Medical School	nd, CF	nd/other
MHH16610	Hannover Medical School	tracheobronchial secretion	respiratory tract
MHH16727	Hannover Medical School	wound swab	wound infection

Table S2: Assembly, annotation and pan-genome information of clinical *P. aeruginosa* isolates.

ID	Total reads	Coverage	Scaffolds	Genes
CH2427	1036963	79	45	6447
CH2500	1060019	98	63	6681
CH2502	976922	62	47	6110
CH2522	1023784	50	32	6095
CH2527	1043430	45	20	6304
CH2543	897689	68	42	6429
CH2560	778901	81	44	6466
CH2563	750638	207	156	6665
CH2582	981196	72	43	6282
CH2591	658419	79	40	6465
CH2597	1051957	42	21	5872
CH2598	796503	34	34	5959
CH2608	953576	62	39	6219
CH2623	746516	82	41	6520
CH2639	898447	86	54	6462
CH2657	888335	76	51	6456
CH2658	1052858	69	36	6276
CH2660	783579	31	17	5753
CH2665	1042969	129	96	6269
CH2672	857588	73	37	6550
CH2674	826381	80	53	6431
CH2675	943818	90	51	6555
CH2677	992389	152	126	6660
CH2678	963297	41	16	5791
CH2680	850043	50	26	5783
CH2682	717428	50	19	6475
CH2685	931734	42	29	6351
CH2687	1165193	55	24	6434
CH2690	929232	79	39	6512
CH2698	903510	59	32	6432
CH2699	1034026	84	46	6548

CH2705	1240819	64	39	6246
CH2706	1092913	71	33	6643
CH2707	1024775	58	25	6717
CH2713	1015084	57	26	6136
CH2718	933175	89	60	6831
CH2724	1039440	45	19	5915
CH2730	1093250	65	35	6102
CH2734	779946	146	91	6283
CH2735	647755	55	36	6032
CH2747	655841	55	26	6509
CH2748	646034	84	45	6534
CH2756	1143624	105	52	6170
CH2764	781513	56	56	5961
CH2804	788368	26	14	5840
CH2824	772900	74	34	6048
CH2829	821886	52	32	5854
CH2848	679884	31	19	5738
CH2860	1129714	35	19	5779
CH2875	796938	48	28	6762
CH2922	1092747	79	41	6320
CH2939	512812	98	48	5895
CH3173	974733	83	53	6429
CH3177	802681	47	33	6007
CH3290	848208	52	37	6157
CH3325	1083130	68	52	6188
CH3462	957111	57	30	6322
CH3466	1223274	60	29	6554
CH3484	10545186	231	192	6365
CH3549	1050132	81	43	6033
CH3570	1351794	98	54	6651
CH3606	1246929	77	46	6391
CH3613	1020076	89	43	6793
CH3648	748725	86	57	6571
CH3797	734237	48	19	6263
CH3882	1019670	57	17	6482
CH3906	1298307	51	31	6154
CH4035	1098198	79	43	6460
CH4083	1000898	52	24	5868
CH4878	559992	44	28	5808
CH4992	565794	79	36	6361
CH5022	595074	108	50	6506
CH5052	587673	80	46	6465
CH5066	620011	42	14	5879
CH5159	555379	103	54	6491
CH5174	536468	92	51	6357
CH5182	614054	32	19	5798
MHH14911	1329007	66	41	6085
MHH14929	735441	59	32	6287

MHH14990	578337	46	19	5954
MHH15015	1402991	86	43	6699
MHH15103	1074926	65	40	6091
MHH15151	1449565	36	21	5692
MHH15187	1175454	37	13	6287
MHH15204	1362002	51	29	5912
MHH15280	1077770	59	34	6167
MHH15286	1137215	131	94	6557
MHH15760	1503676	82	28	6719
MHH15817	1311840	34	17	6549
MHH16050	1645685	41	20	5803
MHH16208	881404	66	40	6320
MHH16371	850446	42	19	5931
MHH16427	754713	61	36	6305
MHH16459	910907	71	57	5869
MHH16513	831770	32	16	6042
MHH16530	1100739	68	40	6363
MHH16563	873572	73	44	6374
MHH16610	765657	31	18	6043
MHH16727	1062340	66	48	6300
Total genes				631,910
Pan-genome				18,319
Unique genes (singletons)				5,539
Core-genome				3,814
Accessory genes				8,966

Table S3: List of fully sequenced genomes in *P. aeruginosa* as on February 15, 2016

Accession ID	**P. aeruginosa strain**
AE004091.2	PAO1
CP000438.1	UCBPP-PA14
CP000744.1	PA7
FM209186.1	LESB58
CP013680.1	AES-1R
CP002496.1	M18
CP003149.1	DK2
AP012280.1	NCGM2.S1
CP004061.1	B136-33
CP006245.1	RP73
CP008739.2	VRFPA04
CP013993.1	DHS01

CP004054.2	PA1
CP004055.1	PA1R
CP006853	MTB-1
CP006937.1	LES431
CP006931.1	SCV20265
CP007147.1	YL84
CP007399.1	F22031
AP014646	NCGM-1984
AP014622.1	NCGM-1900
CP010555.1	FRD1
CP011317.1	Carb01-63
CP012001.1	DSM-50071
CP012066	F9676
LN871187	PAO1_Orsay
CP012679.1	PA1RG
CP013245	VA-134
LN831024.1	NCTC-10332
CP013144.1	Cu1510
CP013696	12-4-4-59
CP008865.1	S86968
CP008866.1	T38079
CP008867.1	T52373
CP008868.1	T63266
CP008869.1	W16407
CP008870.1	W36662
CP008871.1	W45909
CP008864.1	W60856
CP008856.1	F23197
CP008857.1	F30658
CP008859.1	H5708
CP008860.1	H27930
CP008861.1	H47921
CP008862.1	M1608
CP008863.1	M37351
CP013989.1	USDA-ARS-USMARC-41639
CP008872.1	X78812
CP008873.1	F9670
AP014651.1	NCGM257
AP014839.2	8380
AP017302.1	IOMTU 133

Acknowledgements

Acknowledgements

I would like to thank my supervisor and Prof. Susanne Häußler for the opportunity to do my PhD in her group. She provided an interesting topic to work with, and excellent guidance as well as training to progress steadily in that work. She has taught me, how to see the big picture, how to look for interesting questions and how to answer them thoroughly. I appreciate all her contributions of time, perpetual motivation, never ending ideas, advices, and funding to make my PhD experience productive and stimulating. The enthusiasm she has towards her work is contagious and motivational for me. I specially want to thank for her will that made possible for me, to attend many conferences and course works, to communicate with leading scientists. I also thank her for accommodating my "temporal" disabilities and for being accessible almost all the time in spite of her busy schedule.

I would like to thank my mentor Prof. Karsten Hiller and Prof. Michael Steinert for chairing the board of examiners.

I thank the members of my thesis advisory committee, Prof. Dietmar Schomburg and Prof. Irene Wagner-Döbler for their support, stimulating discussions, constructive criticism and for always being present in the TAC meetings amidst their busy schedules.

I am very thankful to Andreas Dötsch, Denitsa Eckweiler and Juliane Düvel who all mentored me during the early days of PhD, taught me how to develop skills towards our group research focus.

My thanks to Piotr Bielecki for the wonderful working experience on the *Escherichia coli* work and also Sebastian Bruchmann for the fruitful experience on *Klebsiella pneumoniae* work.

I thank all the current and past members of the MOBA at HZI and at Twincore for the support and great atmosphere throughout my PhD, especially through the internal symposia. I am very fortunate to have joined a group of caring and lovable colleagues. I thank Monika Schniederjans, Sebastian Bruchmann, Ariane Khaledi and Sarah Pohl for their continuous advices and tips throughout my PhD. Special thanks to Agata Bielecka, Tanja Nicolai, Anja Kobold, Agnes Nielsen and Adrian Kordes for the implementation of Illumina library preparation.

I would like to thank Denitsa Eckweiler, Klaus Hornischer, Sarah Pohl and Matthias Preuße for the excellent bioinformatics and technical support. I also thank Juliane Hartlich for her support during my acknowledgement test. Special thanks to Ariane Khaledi and Silvia Schinner for translating the abstract of this thesis.

Many thanks to Anja Loose, Sabine Schiller and Katharina Hanke for not only taking care of all the administrative work but also for providing a friendly atmosphere to concentrate on work.

I thank all the cooperation partners who kindly provided us with clinical samples. I also thank the Genome Analysis Group at HZI for performing the sequencing.

Thanks to HZI community and Grad School for the amazing and creative atmosphere that provides a great platform to work in and for funding my PhD.

Special thanks to each and every member of SV-Stöckheim badminton club, especially my gang Ariane Passoke, Christian Schütz, Julia Kirchhoff, Martin Jentzsch, Roger König and Sven Marheineke. Special thanks to Karsten Bergmann, Tobias von Chrzanowski and Joachim Wissel for late night games and beers. This club likewise its members played really a major role for refreshing my energy once a week with games, fun and laughs. Thanks to Bishnu and Vinay for introducing me to the wonderful badminton club in Stöckheim.

I also thank lot of friends from Braunschweig for sharing good food, movies, traveling, joyful conversations and especially for bringing color and diversity to my life.

Special thanks to Prof. K. Sekar, my advisor during the research fellow at the Indian Institute of Sciences, introduced me to the scientific world and taught me to be inquisitive. Without his guidance, life wouldn’t be what it is!

Finally I thank my family for all the love, care, support and encouragement they have given me throughout my life. Without which I would not have been able to pursue this research.

There are many others who have influenced my life and career in so many different ways. Although I have not listed every one of you here, know that I am very grateful to each one of you.

Curriculum vitae

Uthayakumar Muthukumarasamy
A2.54, MOBA (Molecular Bacteriology)
Helmholtz Center for Infectious Research
Inhoffenstrasse 7
38124 Braunschweig
Germany
+49 531 6181 3049

E-mail**:** ukm12@helmholtz-hzi.de (or)
Uthayakumar.Muthukumarasamy@helmholtz-hzi.de

Experience

PhD student in Prof. Susanne Häußler group at Helmholtz Center for Infection Research (HZI), Braunschweig, Germany (Sep/10/2012 to present)

Research Assistant (Dec/04/2009 to Jul/31/2012)
As a research assistant under the guidance of Professor K. Sekar, Bioinformatics centre, Indian Institute of Science (IISc), Bangalore, India

Teaching Assistantship (Jun/01/2009 to Dec/01/2009)
Worked as a lecturer in EGS Pillay college of Arts and Science, affiliated to Bharathidasan University, Tamilnadu, India

Education

Master of Science (M.Sc) in Bioinformatics
(A five year integrated postgraduate program, 2004-2009)
Annamalai University, Chidambaram, Tamilnadu, India
CGPA: 8.08/10.0

Research interests

NGS data analysis: genomics, transcriptomics, meta-omics, bacterial adaptation and pathogenesis, adaptive evolution

Computational Structural Biology: sequence structure relationship, molecular dynamics, modeling, docking, drug designing

Skills

Perl, Linux, R, cgi/html, MySQL, data mining and comfortable working in Linux, Windows and Macintosh platforms

Publications (selected papers)

- BACTOME - a reference database to explore the sequence-and gene expression-variation landscape of *Pseudomonas aeruginosa* clinical isolates
  Hornischer, K., Khaledi, A., Pohl, S., Schniederjans, M., Pezoldt, L., Casilag, F., Muthukumarasamy, U., et al.
  Nucleic acids research (2018). gky895, 2018 Oct 1.

(* co-first authors)

- Deep transcriptome profiling of clinical *Klebsiella pneumoniae* isolates reveals strain and sequence type-specific adaptation
  Bruchmann, S., **Muthukumarasamy, U**., et al.
  Environmental Microbiology (2015). 17(11):4690-4710

- In Vivo mRNA Profiling of Uropathogenic *Escherichia coli* from Diverse Phylogroups Reveals Common and Group-Specific Gene Expression Profiles
  Bielecki, P*., **Muthukumarasamy, U***., et al.
  mBio (2014). 5(4), e01075-14

- Homopeptide Repeats: Implications in protein structure, function and evolution
  M. Uthayakumar*, B. Benazir* et al.
  Genomics, Proteomics and Bioinformatics (2012). 10(4), 217-225

- Differential Activities of the Two Closely Related Withanolides, Withaferin-A and Withanone: Bioinformatics and Experimental Evidences
  Vaishnavi, K., Saxena, N., Shah, N., Singh, R., Manjunath, K., **Uthayakumar, M.**, et al.
  PLoS ONE (2012). 7(9): e44419

Curricular activities (selected)

- EMBO Practical Course on Computational Biology: From genomes to systems, Okinawa, Japan, 17/04/2015 - 22/04/2015
- 5th International Human Microbiome Congress, IHMC, Luxembourg, 31/03/2015 - 02/04/2015
- The Goettingen spirit summer school in "Biological Research in Dentistry" Sep/28/2014 - Oct/3/2014, Goettingen, Germany
- EMBO-GELC High-throughput next generation sequencing applied to infectious diseases, Sep/15/2104 - Sep/25/2014 at Institut Pasteur de Tunis, Tunisia
- RHCE certification, Red Hat Certified Engineer (Linux Certification from RedHat)

Other

Date of Birth	May/25/1987
Sex	Male
Languages	English, Tamil
Nationality	Indian
Hobby	Badminton, Traveling

Declaration I hereby declare that all the information given above are true to the best of my knowledge and nobility